Responsibility for Self: Agency and the Attitudes

By

Mark Rosner

A thesis submitted to the Graduate Program in Philosophy

in conformity with the requirements for the degree of

Doctor of Philosophy

Queen's University

Kingston, Ontario, Canada

August 2015

Abstract

This thesis defends the claim that the core idea of moral responsibility is fixed by our best theory of agency. Such a theory concerns the proper conditions of attribution of an attitude or an action to an agent for the purposes of moral appraisal – or what I call the rational relations view. The first chapter attempts to outline and motivate the conception of agency that has as its core the idea that agents are responsible for the judgement-sensitive attitudes that can be appropriately attributed to them and that they can be morally responsible for those attitudes when their judgements have morally objectionable contents.

In my second chapter I present and argue against what I take to be the strongest alternative to the rational relations view, the theory of moral responsibility that has been elaborated by Harry Frankfurt over a number of years. The third chapter addresses a concern that conceptions of responsibility that are too closely tied to theories of agency can either be too superficial in their assessment and evaluation of the agent or actually unfair in their determinations of responsibility.

The fourth chapter extends this discussion of the unfairness charge by explicitly addressing the question of the value of moral responsibility. My final chapter concerns an issue that lies at the intersection of questions in the philosophy of agency and moral responsibility: how to make sense of and be open to criticism for our acts of irrationality. I take the example of akratic, or weak-willed action, where an agent acts contrary to her better judgement (or what she judges best in a situation) as a paradigm instance of irrationality. I argue that RR has adequate resources to make sense of this phenomenon.

At base, this thesis aspires to show that by elaborating an attractive picture of our agency we can at the same time shed light on what it means to be a responsible agent – one for whom it is intelligible and valuable to say we are morally responsible for our attitudes and actions.

Acknowledgements

I would first like to acknowledge my gratitude to my supervisor Rahul Kumar. It was through coursework with him that my interest in this area of philosophy was first sparked and through later conversations that my views were formed. His criticism, patience and guidance have greatly improved my own philosophical understanding of the topic, as well as the thesis, and helped me see it to completion.

I would also like to acknowledge a large debt to David Bakhurst and to Stephen Leighton. My coursework and conversations with them during my time at Queen's greatly expanded my understanding of what philosophy is and how it can be done well. I owe David Bakhurst a special debt for the insightful and helpful comments he provided for my thesis.

I would like to acknowledge the many friends and colleagues – both inside the Department of Philosophy and in the broader Queen's community – I made during my years in Kingston. Helpful advice can take many forms. Their friendship and camaraderie helped me in many ways to develop both as a philosopher and as an individual.

This project was aided by the generous support of the Department of Philosophy at Queen's University as well as the Social Sciences and Humanities Research Council of Canada. I am grateful to them both.

Finally, I would like to thank my family: my parents, Harriet and Cecil, my brother, David and my sister, Michelle. Their encouragement and interest in my work helped sustain me over the course of this project. To my son Ben and my partner Karelle Arbez I thank you both for your love and unwavering support. To Karelle: I owe you much more than I can properly say.

Table of Contents

Responsibility for Self: Agency and the Attitudes

Chapter 1- Introduction

This thesis defends the claim that the core idea of moral responsibility is fixed by our best theory of agency. Such a theory concerns the proper conditions of attribution of an attitude or an action to an agent for the purposes of moral appraisal – what I call, following other authors, the rational relations view. Thus moral responsibility, as a concept, is best understood in terms of the theory of agency outlined in the following pages.

This presupposes a division of labour in outlining a theory of moral responsibility. Theories of responsibility will not, on my view, determine what in fact we are responsible for. That is, they will not say, in any direct fashion, when we should bear the cost of our foolish actions or when should we benefit from our fortunate circumstances. Determining what the substantive moral norms are – the prohibitions and requirements that would structure our deliberations and guide our actions – is the job of normative moral theorizing, rather than a theory of responsibility. A normative moral theory will outline and, to some extent, codify these norms into a structured theory. What theories of responsibility do, however, is to determine when it is correct to say that agent 'x' has failed to live up to her responsibilities or when it is true agent 'y' is not to be found responsible even though she violated a moral norm. Theories of responsibility set out the appropriate conditions of moral appraisal.

This means that a good theory of responsibility will presuppose a substantive normative moral theory in determining when an agent is responsible. What this theory looks like is not a question that I am able to answer in this thesis. What I do hope is that

the theory of responsibility outlined here is, as much as possible, agnostic between competing moral theories. So, supposing one were a committed rule-utilitarian, such a position would be consistent with the theory of moral responsibility outlined here. Additionally, if one held that a version of contractualism was the best moral theory available this would also be consistent with the theory of moral responsibility outlined here. While intuitions that flow from our substantive moral theory will colour our thoughts as to who or what is responsible for something, I claim that whether an agent is an appropriate target of such a thought can be settled without requiring us to antecedently align ourselves with a favored moral theory.

On my view, the concept of responsibility is most properly connected with the concepts of ascription and appraisal (or the proper conditions of appraisal). The idea is not novel; Joseph Raz has claimed that the core idea of responsibility is one of ascription and its relation to the conditions of appraisal. He writes that the "theory of ascription is concerned with the conditions in which blame or guilt can be ascribed to people…The theory of ascription deals with the ascription of blame and praise to people who fulfilled or failed to fulfil their responsibilities" (Raz 1990: 12). What is distinctive regarding my claim regarding the nature of moral responsibility is that I argue that the conditions of ascription and appraisal are tied to the presence or absence of a judgement-sensitive attitude in an agent who is governed by an ideal of rationality. This conception of responsibility is thus dependent on a theory of agency which outlines the proper conditions of attribution of attitudes to agents for purposes of moral appraisal.

The conception of responsible agency that is in the background of the theory of moral responsibility is one that can be termed 'responsibility as attributability'. This view

can be summed up in the formula: '(RA) An attitude or action is attributable to a person for purposes of responsibility if and only if it is connected to the agent in such a way that it can serve as (part of) a basis for appraisal'. This formulation, while general, has two virtues. First, it clearly shows the need for a developed theory of agency when we are making determinations of responsibility. We must be able to say how and in what way an attitude or an action is relevantly connected to an agent in order for a judgement of responsibility to be applicable. To anticipate, the view that I defend holds that the relevant connection between an agent and her attitudes and action consists in a normative relation. This normative relation is what an agent sees as good (but not necessarily best). In other words, it is a reflection of her judgemental activity, as an agent who is governed by an ideal of rationality, broadly construed. While this is clearly schematic it shows a way of formulating the terms of adequacy for a theory of agency; making use of such a framework can help us determine what actually is the best theory of moral responsibility.

Second, this formulation allows us to understand certain moral issues which may otherwise escape consideration when we approach the theory of responsibility from a different direction.[1] Thomas E. Hill Jr. writes regarding the vice of social snobbery that

> contemporary ethics is mostly focused upon right and wrong *action*, duty and supererogation, or rights and welfare, but our questions about snobbery belong to the ethics of *attitudes*. Too often it is assumed that objectionable attitudes are simply those that lead to wrong *actions*, but in some cases, such as snobbery, this seems to have matters backwards…snobbish acts are wrong, in part, because of the objectionable attitudes they manifest, not the reverse (Hill 1991a: 156 – emphasis in original).

Whether or not we accept Hill's Kantian inspired claim that the vice of snobbery is best understood as a failure to acknowledge the basic worth of all humans as persons, I think it is plausible that there are important ethical questions that turn on what Hill calls

[1] I canvass and criticize different approaches to moral responsibility in chapters two and three.

an ‘ethics of attitudes’. A theory of moral responsibility will, as a result, be required to be able to explain not only how an agent can be responsible for her wrong actions but also how she can be criticized for objectionable attitudes. RA’s broad focus on the proper conditions of ascription of attitudes for the purposes of appraisal shows how an ethical theory that incorporates an ethics of attitudes can attribute responsibility to agents for their objectionable attitudes. While it is not necessary that an ethical theory do so, I think the fact that RA can accommodate such an extension of ethical theory is a point in its favour.

Support for this conception of agency also comes from a different direction. It is difficult to over-emphasize the massive and profound impact that Peter Strawson’s article “Freedom and Resentment” has had on the free will and moral responsibility debate. It has spawned new approaches to the topics, specifically ones that emphasize an agent’s quality of will or her reactive attitudes, while shunting others to the sidelines, in particular attempts to ground free action in an explicitly metaphysical account. For my purposes, I think there are two points of Strawsonian inspiration that are captured by the approach to moral responsibility outlined here.

First, RA emphasizes, as Strawson was fundamentally interested in doing, the quality of will of the morally responsible agent. As he took pains to underline, people care quite deeply and pervasively about the quality of will that others show towards us – whether that will expresses good will or regard or rather is an expression of hatred or contempt. RA puts front and centre the attitude or action that an agent held or performed. We determine whether or not an agent is in fact responsible by attending to that quality of will. Thus, a focus on the quality of will of the agent is a pre-condition for making a

determination of responsibility according to RA. Further, if in fact RA is consistent with an ethics of attitudes, then I think this is another reason for thinking that RA properly emphasizes the importance of the quality of will of the responsible agent.

A second point of Strawsonian inspiration and support for the approach made use of in the thesis comes from the emphasis and importance Strawson placed on interpersonal relationships in helping us to understand the conditions of moral responsibility. For my purposes, I see in the idea that interpersonal relationships are key to understanding moral responsibility the claim, most fundamentally, that the *value* of moral responsibility lies in the fact that it allows us to participate in important worthwhile relationships with other persons. What I argue for in the thesis is that a proper account of moral responsibility must shed light on the value of moral responsibility as well as on its conditions of application. I claim we can do so by attending to the interpersonal significance of the attaining the status of being a morally responsible agent – this is what is valuable about being a morally responsible agent. So facts about our ability to stand in meaningful relationships determine the value of being a responsible agent on my view.

This is, of course, not the only possible interpretation of Strawson's article (nor is it exhaustive as I do not make use of the idea of the reactive attitudes in any extended way in the thesis). For example, I do not interpret Strawson's emphasis on our propensity to react with, say, resentment to the ill-will displayed by another agent as evidence that we should determine who is responsible by looking to who we are prepared to hold responsible. That said, I do think there is a recognizable sense in which the approach to

thinking about agency and moral responsibility made use of in the thesis builds on insights Strawson first put to the fore.[2]

The thesis divides itself into five chapters. In the first chapter, I outline how it is that I understand the relationship between RA and the more specific version of responsible agency that I favor RR. More generally, the first chapter attempts to outline and motivate a conception of agency that has as its core the idea that agents are responsible for the judgement-sensitive attitudes that can be appropriately attributed to them and that they can be morally responsible for those attitudes when their judgements have morally objectionable contents. Following T.M. Scanlon and Angela Smith I hold that it is our activity as agents who can recognize and be governed by reasons that determines when we are in fact responsible for our attitudes and our actions. I then proceed to defend this conception of responsible agency against a number of objections. In particular, I note that we can make distinctions within RR, between wide and narrow versions of the theory and between thick and thin versions of the theory, distinctions which, once noted, can at once defuse some of the concerns that have been raised against RR and at the same time display the resources RR has on offer as a theory of responsible agency.

RR is a particular interpretation of the general formulation outlined by RA. I recognize that it is not the only possible interpretation of this general formula. As a result, in the second chapter, I present and argue against what I take to be the strongest alternative interpretation of RA, the theory of moral responsibility that has been elaborated by Harry Frankfurt over a number of years. Frankfurt's early work on the

[2] I think a useful contrast to the way I invoke Strawson can be made by looking at Michael McKenna's recent work on the nature of moral responsibility, *Conversation and Responsibility*. See Mckenna (2012).

hierarchical conception of agency set the stage for subsequent discussions of agency. His more recent work on the importance of caring and love has also opened new ground in discussions regarding the fundamental aspects of our agency and the understanding of terms such as 'activity' and 'alienation'. While I recognize that I bear a debt to Frankfurt and that there are significant insights to be gained from attending to his work, I ultimately argue that both his early and more recent views are open to objection by failing to give proper place to the role that reason plays in the life of a morally responsible agent. RR, by contrast, gives proper place to the role that reasons play in our psychology and I show that as a result it is to be preferred to Frankfurt's account.

The third chapter addresses a concern that conceptions of responsibility that are too closely tied to theories of agency can either be too superficial in their assessment and evaluation of the agent or actually unfair in their determinations of responsibility. In a way, this chapter attempts to defend the thought that the conception of moral responsibility that flows from the conception of agency outlined and defended in chapter one is not only a legitimate specification of the concept of moral responsibility but that it is in fact preferable to views of responsibility that reject the centrality of RA. Drawing on the theory of agency defended earlier, I claim that contrary to a common view, RR is not a superficial take on who the agent is and so constitutes a legitimate basis for moral appraisal of the agent. Further I argue that the charge of unfairness can be rebutted in a direct and an indirect fashion. The direct rebuttal disputes the background claim of the unfairness strategy – that we should determine who is responsible by looking to who we are prepared to hold (fairly) responsible. The indirect strategy claims that proponents of the unfairness charge have overreached. That is, the moral theorizing that led them to

suggest that conceptions of responsibility like RA are problematic is wrong because there are important and relevant moral arguments that actually push us to adopt conceptions of responsibility that place attributability front and centre.

The fourth chapter extends this discussion of the unfairness charge by explicitly taking on the question of the value of moral responsibility. I start by criticizing the best available compatibilist account of the value of moral responsibility that is due to John Martin Fischer. While Fischer's use of the idea of a dramatic narrative (or 'story') as a device to shed light on our self-expressive activity (and so to illuminate the value of responsibility) is innovative, I find it lacking. On the one hand, I hold that over-emphasis on the idea of a narrative can obscure cases where, in the absence of a dramatic narrative, we nonetheless are morally responsible for an action and where such exercises of our agency are valuable. On the other hand, I claim that Fischer's account pays insufficient attention to the interpersonal element in the value of moral responsibility. I thus supplement his account by drawing on resources available within RR. In particular, I claim that RR's emphasis on the ideal of answerability is a way of properly acknowledging the interpersonal aspect of the value of moral responsibility. The fact that we are answerable for our agency – what gives content to the phrase 'responsible agency' – is what connects us with, and makes sense of, the various relationships we find ourselves participating in. Thus I hold that RR can not only offer an attractive theory of agency but can make good sense of why we would value such an account.

My final chapter concerns an issue that lies at the intersection of questions in the philosophy of agency and moral responsibility: how to make sense of and be open to criticism for our acts of irrationality. I take the example of akratic, or weak-willed action,

where an agent acts contrary to her better judgement (or what she judges best in a situation) as a paradigm instance of irrationality. I argue that RR has adequate resources to make sense of this phenomenon. In particular, I claim that the primacy RR places on our judgemental activity is a virtue of the account. I do so by criticizing some recent arguments by Nomy Arpaly and Alison MacIntyre that it can be, on occasion, rational to act contrary to one's own best judgement. I claim that that the phenomena that they draw attention to, specifically cases of 'inverse akrasia' as Arpaly terms them and instances of action that subjectively appears to be irrational but that, from the agent's own perspective would appear to be the thing to do, can be made sense of by RR while still preserving the intuitive thought that an agent, in acting against her best or better judgement displays a basic form of irrationality. That is, RR can explain how an agent can be responsible for her irrational action, all the while preserving the thought that it is, at base, irrational. Further, I claim that RR need not be revised in any serious way in order to accommodate this thought. R. Jay Wallace has argued that the only way to make sense of the phenomenon of akratic action, and in particular, strong-willed action in the face of temptation, is by positing some type of basic executive capacity for self-determination on the part of the agent. However, I again draw on the fact that RR holds that all of our judgemental activity (including judgements that we nonetheless do not, on balance, rationally endorse) are reflective, in some way, of our agency. In this fashion I try and show that RR, as a theory of agency, need not require supplementation by some further volitional capacity in order to make sense of weak- and strong-willed action.

At base, this thesis aspires to show that by elaborating on an attractive picture of our agency we can at the same time shed light on what it means to be a responsible agent

– one for whom it is intelligible and valuable to say we are morally responsible for our attitudes and actions. It is an attempt to show how our agency and activity as beings that can recognize and respond to reasons informs our practices of attributions of responsibility, explains the value we find in being, and being held, morally responsible and structures our attitudes towards common failings we have as beings with rational powers.

Chapter Two: The Rational Relations View: Wide and Narrow, Thick and Thin

Section 2.1 – Responsibility, Agency and Attributability

When is an agent morally responsible for her attitudes and actions? This is the question I intend to answer in this chapter by sketching an account of responsible agency, that I follow Angela Smith in labelling the rational relations view.[3] Such an account seeks to provide the necessary and sufficient conditions for determining whether or not an agent is morally responsible for an attitude or an action. While this investigation does not imply anything regarding what moral assessment of the agent is appropriate, it will provide the necessary guidance to determine whether the agent is a credible candidate for being responsible, and whether there is some kind of defeater that might limit whether we consider an agent responsible in the circumstances. So, in sketching the necessary requirements for being morally responsible, the rational relations view will offer resources to understand why small children, for example, are not normally considered responsible for their (objectionable or praiseworthy) attitudes or actions, but competent adults are. These necessary conditions will also show how certain excuses are meant to function, diminishing or absolving agents of responsibility because of coercion or ignorance. In sum, the rational relations view is a systematic theory of responsibility that spells out the necessary and sufficient conditions on morally responsible agency.

First, some comment on the concept of 'moral responsibility' is needed. Philosophers frequently find the account of responsibility put forward by P.F. Strawson in his classic essay "Freedom and Resentment" as providing the contours and conditions of moral responsibility, defining what is at issue in the debate over what responsibility is. Whether or not they agree with Strawson's arguments in that essay, Strawson's account

[3] See Smith (2005).

supplies 'neutral ground', if you will, providing the data with which philosophers must work in order to come to an adequate account of the nature of moral responsibility. Outlining this 'raw material' will be useful as it will show how the view I outline below links up with the wider literature on the nature of moral responsibility, and how it embodies an attractive account of the nature of moral responsibility.

Strawson's account begins by noting what he calls some 'commonplaces'. He writes:

> The central commonplace that I want to insist on is the very great importance that we attach to the attitudes and intentions towards us of other human beings, and the great extent to which our personal feelings and reactions depend upon, or involve, our beliefs about these attitudes and intentions. (Strawson 1962: 75)

The qualities of other agents' wills, or what their attitudes or intentions are, are of fundamental importance to us on Strawson's account. For what quality of will an agent displays determines whether she is showing goodwill or affection, and so may be liable to a typically positive response from the object of those attitudes (such as gratitude), or some form of contempt or malevolence, which will legitimately provoke another form of response (such as resentment). Strawson thinks that it is practically and psychologically impossible for someone to deny that we care about the various attitudes and intentions other individuals have towards us.[4] And so, it seems, we ought to care about the responses that other agents are liable to take towards us.[5]

There are two central elements to Strawson's account of moral responsibility. First is the idea that we are part of, or involved with, or participants in, a variety of relationships with a variety of different individuals. Strawson mentions some examples of

[4] Even if what Strawson argues is true, whether one ought to care is a separate question.
[5] For a variety of different interpretations of Strawson's argument, see McKenna and Russell (2008).

the great variety of relationships we share with other people: "as sharers of a common interest; as members of the same family; as colleagues; as friends; as lovers; as chance parties to an enormous range of transactions and encounters" (Strawson 1962: 76). These examples of inter-personal relationships are central to his conception of moral responsibility. These relationships are ubiquitous, fundamentally important and are, arguably, an indispensable "part of the general framework of human life" (Strawson 1962: 83).[6] They are indispensable because they embody what it is we actually care about in human life, they

> emphasize how much we actually mind, how much it matters to us, whether the actions of other people – and particularly *some* other people – reflect attitudes towards us of goodwill, affection, or esteem on the one hand or contempt, indifference, or malevolence on the other (Strawson 1962: 76 - emphasis in original).

When agents form part of these various relationships, as they must on Strawson's view, they tend to display what he calls the "reactive attitudes" (Strawson 1962: 76, italics removed). These are attitudes of resentment or gratitude, contempt or affection that agents hold towards other individuals who are part of a relationship with them. So, if one participant in a relationship expresses a certain attitude (or quality of will) of contempt for another participant in the relationship – say a brother making a particularly biting, mean-spirited comment to a sister – then the other participant will be prone to have certain feelings in response to the attitude expressed. The sister might reasonably complain of her brother's treatment and resent the problematic attitude he expresses in his actions; she even may take further steps to see her brother sanctioned (say by her

[6] While I think the importance of interpersonal relationships to an adequate account of moral responsibility is widely held, I do not claim that Strawson's assertion that we are naturally committed to such relationships and that we cannot put these general web of relationships 'up for review' is widely held.

parents) for his offending conduct. While the content of the reactive attitudes will vary with the particularities of the case, the idea that such attitudes provide some grasp on *who* is morally responsible is what is important here. That is, the reactive attitudes attach to other members of inter-personal relationships. They are only properly held towards those who are, or can be, called participants in a morally significant inter-personal relationship.[7] So if an agent is open, in principle, to the reactive attitudes then we have a grasp on who is a legitimate candidate for a responsibility attribution.

I think the idea that an agent is an apt target of the reactive attitudes is a common and influential understanding of the nature of moral responsibility.[8] If an agent is the proper target of feelings of resentment, given the quality of will that she may have displayed, then she can be said to be a morally responsible agent. Looking to these attitudes and the relationships in which they are embedded provides the way in which we understand who is morally responsible and under what conditions. Thus, it provides an account of responsible agency, similar to the one I offer below.

My strategy will be to base the conception of moral responsibility on the account of responsible agency outlined here such that a unified conception of agency and responsibility will follow. While some may question the basic thrust of this approach I hope to make it clear why I think this strategy sheds on light on the sense of what it means to be morally responsible for our attitudes and actions. [9]

---

[7] The difference between agents who are participants in some form of interpersonal relationship and who can be participants is important, but I don't mean to comment on whether Strawson had any particular interpretation in mind here.

[8] Wallace (1996) is perhaps the most influential recent Strawsonian treatment of moral responsibility; Watson (1987b) is another important treatment of responsibility that takes Strawson seriously.

[9] See, for a differing approach, Wallace (1996). Wallace holds that we should adopt the stance of a 'moral judge' in order to tease out the nature and norms of responsible action. While adopting

The sense of responsibility that the rational relations view presupposes can be termed "*responsibility as attributability*" (Scanlon 1998: 248). According to this view, when we are attempting to determine whether an agent is responsible for some attitude or action, the question that guides this investigation will be: is this action or attitude properly attributable to an agent, or in the relevant sense 'ours', for the purposes of making a moral judgement or assessment of that agent?[10]

Responsibility as attributability (RA) holds that:

> '(RA) An attitude or action is attributable to a person for purposes of moral responsibility if and only if it is connected to the agent in such a way that it can serve as (part of) a basis for moral appraisal'.[11]

RA holds that it is a certain kind of connection between an agent and her attitudes and actions that allows those attitudes and actions to be objects of moral appraisal (whether positive or negative). As it does not specify what kind of connection is required, more than one account of responsible agency is consistent with RA. An implication of RA is that where this connection between an agent and her attitudes and actions is radically attenuated or completely absent, then the corresponding judgements of responsibility will be attenuated or absent. If the absence of this connection is what many authors mean when they describe agents who are 'alienated' from their actions or who have 'disowned' certain of their attitudes and subsequently view them as 'outlaws' in

---

such an interpersonal perspective is important, I do not think it essential to outlining the nature and grounds of responsible action. I say more about this in chapter four.

[10] What makes an attitude 'ours' for the purposes of moral responsibility is a central question in discussions regarding the nature of agency. If it is possible (e.g. in cases of hypnotism) for an action to be intentional but for which an agent should not be considered responsible, then an action can be 'ours' in a sense that is not relevant to attributions of moral responsibility.

[11] Henceforth when I discuss attributability I will deal only with the kind that is relevant to moral responsibility, unless otherwise stated, given that attributability can be used in a variety of different contexts (e.g. in discussions in the philosophy of mind or in cultural/political discussions regarding the nature of identity).

their psychological and practical lives, then RA can make sense of why those agents normally abjure from judging them to be properly responsible for those actions and attitudes. Moreover, once we spell out the nature of the connection that is embedded in RA, we will have made some preliminary progress in determining not only the basic contours of responsible agency (and as such who is to be exempted from consideration as being a responsible agent) but we will also have given some direction as to how to understand certain kinds of excuses that are normally thought to absolve agents from responsibility (such as ignorance, diminished capacity and some more extreme, hypothetical cases, of manipulation). I will try to flag how such excuses are to be understood as I spell out the rational relations view (which I label'RR').

Section 2.2 – Agency and Our Attitudes

The conception of agency mentioned in RA follows a view put forward by Scanlon which claims that each agent is a "conscious, rational, embodied creature" with a "stream of conscious thoughts and experiences…united by a degree of constancy in its elements, by the intentional content of those elements and by its supposed causal basis" (Scanlon 2002: 167).[12] On this picture, there is a certain basic level of unity or coherence in a person's psychic life in both a diachronic and synchronic sense. Over time there is, for the most part, a large degree of continuity between a person's cognitive and affective interests, actions and reactions. What a person believes and holds to be true, what she rejects and is uninterested in, what she enjoys and takes pleasure in and what she judges to be good, as well as what she dislikes and finds to be bad, are, *grosso modo*, fairly

[12] Scanlon also notes that there is some dispute as to the status of the causal basis – for my purposes, this point is irrelevant. For a contrary view, see Sher (2009).

continuous across the life of an individual. This sense of coherence is compatible, however, with alterations and changes in person's psychic life over a period of time.

Moreover, at any one time, it makes sense to say that there is a substantial amount of cross-reference between the various elements of an agent's psychic system. Whether it is recalling the content of a past experience, defending a longstanding belief that one holds or fulfilling a particular intention as part of a larger plan that one is engaged in carrying out, these various attitudes and actions make reference to other aspects of the agent's psychology. While Scanlon does not set out in any detail to what extent the various features of a person's life must be related in order for an agent to have a sufficient degree of continuity, we can stipulate for now that there must be some basic level of connection between the various elements. To anticipate, RR will fill out such details, providing a test whereby we can determine whether something falls within a person's sphere of agency.

Scanlon continues by emphasizing the fact that we are creatures who are rational. The fact that we are rational entails the claim that we are creatures who are capable of "making judgements about reasons and hence having judgement-sensitive attitudes such as belief and intention" (Scanlon 2002:168). On Scanlon's view, agents are capable of making judgements about reasons as well as being moved by those judgements – reasons for belief, for intention and action and reasons for emotion and feeling.

Judgement-sensitive attitudes, on Scanlon's account, involve "a complicated set of dispositions to think and react in specified ways" (Scanlon 1998: 21). What makes these attitudes ones that we can attribute to the agent is that they are sensitive, at least ideally, to our judgements concerning what reasons there are for and against their various

contents and it is part of that attitude that it ought to be sensitive to those judgements. These judgements, though, do not have to be conscious judgements in order to be considered to reflect a person's agential activity. As long as they are properly connected to an agent's rational capacities, regardless of an agent's occurent awareness, RA will count them as being attributable to an agent.

More specifically, these are the "attitudes that an ideally rational person would come to have whenever that person judged there to be sufficient reasons for them and that would, in an ideally rational person, 'extinguish' when that person judged them not to be supported by reasons of the appropriate kind" (Scanlon 1998: 20).

Referencing this ideal allows us to make sense of these attitudes as ones that are generally responsive in the kinds of ways gestured at by the references to coherence and continuity laid out above, to our judgements about reasons. In the ideal case, an agent's attitudes and actions would be completely and unfailingly responsive to her judgements of reasons. So, should such an ideal agent judge that a course of action, all things considered, is to be taken, then she would adopt this attitude and put in into action. Conversely, should such an agent determine that a reason for action is bad she would not make use of such a reason in her subsequent deliberations and would avoid or reject actions that relied for their justification on this claim. Thus, the kind of 'normative pressure' that an agent feels when she fails to keep a longstanding personal commitment to not smoke can be explained by the fact that this agent fails to live up to, as it were, the content of her own judgement. This pressure arises from the fact that the agent recognized the force of the reason in question but did not properly translate such a recognition into her other attitudes and actions.

Thus it is a mistake to think that the various connections a judgement-sensitive attitude will have, rationally speaking, with other elements of an individual's psychology and actions are not relevant to determining whether or not we ought to attribute an attitude to an individual. As Scanlon notes, sometimes such an attribution is justified just because that attitude is that which makes best sense of the individual's thoughts and actions. So, the fact that a person is disposed to help someone in distress is thought to be good evidence that the individual cares (possesses a judgement-sensitive attitude) that we ought to help those in distress when we can or that she cares about her fellows sufficiently to reach out to them in times of great need. Angela Smith formulates this point as a conditional: "if one judges some thing or person to be important or significant in some way, this should (rationally) have an influence on one's tendency to notice factors which pertain to the existence, welfare, or flourishing of that thing or person" (Smith 2005: 244). While fatigue and other factors can lessen the probative value of this evidence by interfering with the normative connection between attitudes and action, the fact that all things being equal this inference from the attitude to the action and from the action to the attitude seems strong lends support to the idea that judgement-sensitive attitudes implicate a broad swath of an individual's psychology and are prime candidates for responsibility-attributions.

Moreover, in the case of some attitudes, such as for example certain forms of romantic or parental love, the connection between a judgement-sensitive attitude and "general patterns of awareness" (Smith 2005: 244, n.14) as well as tendencies and dispositions to act, are tighter. For the tendencies and dispositions seem constitutive of possessing the attitude at all. The having of such tendencies and dispositions, of care and

concern for instance, is what it means, at least in part, to love a person. Exhibiting this range of concerns constitutes the possession of the attitude itself. Conversely, the absence of these general patterns of awareness (being aware, for example, of where your very young baby is located or who they are with) can be grounds for determining that one does not in fact possess a particular judgement-sensitive attitude, in this case parental concern or love.

As a result, calling an individual belief or intention a judgement-sensitive attitude can be misleading insofar as it suggests thinking of this attitude as a kind of capsule, one that can be properly grasped without reference to other elements of an agent's psychology. Scanlon emphasizes, as I am keen to as well, that when an agent has such an attitude they will possess "a complicated *set* of dispositions" (my emphasis) which will manifest itself in a variety of ways in a variety of contexts. There is a holistic element to judgement-sensitive attitudes that implicates other elements of an agent's psychology and demands a broader perspective when we cash out the content of a judgement-sensitive attitude.[13]

Section 2.3 – Responsibility and the Rational Relations View

Given this brief sketch of the nature of agency that is assumed in RA, I want to focus now on the responsibility component. Delineating the nature of the responsibility in

[13] See Davidson (2004). According to Davidson, this phenomenon is the holistic character of the mental, governing all propositional attitudes that an agent may possess (what I take to be equivalent to judgement-sensitive attitudes). He writes: "The meaning of a sentence, the content of a belief or desire, is not an item that can be attached to it in isolation from its fellows. We cannot intelligibly attribute the thought that a piece of ice is melting to someone who does not have any true beliefs about the nature of ice, its physical properties connected with water, cold, solidity, and so forth. The one attribution rests on the supposition of many more – endlessly more" (p.184).

question can elucidate how we can make the theoretical move from RA to RR.[14] According to Smith, "to say that an agent is morally responsible for some thing, on this view [RR], is to say that that thing reflects her rational judgment in a way that makes it appropriate, in principle, to ask her to defend or justify it" (Smith 2008: 369).[15] We are answerable for those things that 'reflect' our rational judgement just because those things are, standardly, judgement-sensitive attitudes. It is a datum of RA that our judgement-sensitive attitudes are an immediate expression of what we take to be the case or what we take to be of significance. While we may not be able to formulate adequate (or even intelligible) reasons for why we hold the attitude in question or why we performed the action in question, if this justificatory request is in good order, then the object of such a request is a candidate for the kind of responsibility that we are after.

If an agent makes a controversial claim, say, that a certain kind of classical music is the best form of musical expression, this seems to reflect what is of significance in the agent's eyes with respect to that artistic domain. What's more, it seems appropriate for another agent (or for the agent herself) to be able to ask her to justify or defend this claim: 'What makes this type of music the best?' 'How did you come to this determination?' seem to be appropriate kinds of questions even if there is no answer in

---

[14] To anticipate, RR holds: 'There is a normative connection between attitudes and actions and our capacity for evaluative judgment such that the presence of this connection determines if and only if we are open to moral appraisal and assessment for our attitudes and actions'.

[15] While I am wary of proliferating labels, I think the best way to understand the core thought behind the rational relations view is to see its conception of responsibility as a kind of *answerability*. This is perhaps not an uncontentious label for Gary Watson has made much use of the term and would not, I believe, accept RR as it is outlined here. For his bifurcated account of the nature of responsibility, see Watson (1996). Nonetheless, it seems to me to be fitting if we understand answerability not in terms of a present (past or future) ability or even propensity to answer a justificatory request, but rather the simple capacity (or capacities) to do so, as capturing the basic sense of responsibility that Smith and Scanlon are after. That is, they seek the conditions that would provide an answer to the following question: 'is this the kind of being to whom a justificatory request is appropriately addressed for the purposes of moral appraisal'?

the offing. Whether we agree or disagree with the judgement regarding musical worth (or find much to criticize) it seems that the judgement expresses what an agent takes to be of significance in her situation and is one that is open, in principle, to requests for justification (or modification or even withdrawal depending on how the subsequent discussion proceeds).

Why is it, though, that we can request a justification for this type of attitude but not, for example, for a person experiencing a severe migraine? Elsewhere, Smith formulates the conception of responsibility as follows: "In order for a creature to be responsible for an attitude, on the rational relations view, it must be the kind of state that is open, in principle, to revision or modification through that creature's own processes of rational reflection" (Smith 2005: 256). The kind of state that Smith has in mind is one captured by judgement-sensitive attitudes.

What it is about this state that makes it a candidate for a responsibility attribution is the content of the state itself. It is the content of a judgement-sensitive attitude that is sensitive, in principle, to an agent's judgements regarding various reasons. Given that we possess states with these kinds of content, and given that they can be modified by an agent's own rational capacities, they seem to be legitimate candidates for the kind of responsibility outlined above.

Fully stated, RR holds: 'There is a normative connection between attitudes and actions and our capacity for evaluative judgment such that the presence of this connection determines if and only if we are open to moral appraisal and assessment for our attitudes and actions'.[16] RR thus offers a unified account of agency and responsibility.

[16] What RR claims is that what determines our status as an agent – the capacity for evaluative judgement – is also what determines whether we are open to moral appraisal. The thought is that

Section 2.4 – Rational Relations View Wide and Narrow, Thick and Thin

In the rest of this chapter, I want to take up the challenge of defending RR against a series of recent criticisms that aim to undermine its ability to account for practices of moral responsibility. I will respond to a series of charges that have been levelled at RR with the aim of showing how it can not only rebut these challenges but display the resources to offer as a systematic theory of moral agency and responsibility. In particular, I hold that each charge ultimately misses its target because we can distinguish a number of different versions of RR: wide versus narrow, thick versus thin. The charges fail, I will argue, because while they are plausible against a narrow and thick version of RR, the wide and thin version of RR is the best articulation of the view and can avoid the criticisms offered.

I'll proceed by first considering the objections to RR that potentially trade on the distinction between a wide and narrow reading of RR, elaborating the distinction in the process. I'll then consider the distinction between thin and thick versions of RR and the relative advantages of the former.

Wide RR holds that we must make essential reference to other psychological states of an agent in fixing the content of a person's judgement-sensitive attitude. Narrow RR, in contrast, holds that we need not make reference to other elements of a person's

---

moral appraisal of an action is primarily a matter of recognizing and responding to the content embodied in an intentional action, and this content is only open to appraisal if it is normatively connected to an agent's rational capacities. This is not a trivial claim regarding the nature of responsibility. Alternative views hold, for example, that an agent is only responsible if, and only if, the attitude in question is under her 'control' (where control is understood to imply conscious awareness of the attitude and to imply the ability to change). Further, Strawsonian views claim that if we are apt targets of the reactive attitudes then this outlines the conditions of responsible agency. But what renders us apt targets need not be a normative connection between our attitudes and our rational capacities. This is not to argue for RR, but to show how it relates to alternative views elsewhere defended.

psychology, or their other judgement-sensitive attitudes, in fixing the content of our judgements.

More particularly, narrow RR sees our evaluative judgements as similar to singular conscious acts of judgement – and that the content and importance of such judgements can be fixed without meaningful reference to the whole array of other judgements that individuals hold. The narrow interpretation holds that it is not a precondition of moral assessment that we refer to the various other judgement-sensitive attitudes an agent possesses in order to fix the content of an agent's attitude.

The wide interpretation of the rational relations view sees the content of our evaluative judgements as both reflected in our explicit judgements as well as fixed by the other judgements that a person holds. That is, the content of a judgement for which we are to hold an agent accountable is fixed not only by her explicit pronouncements but also by other aspects of her psychology (not necessarily consciously held) which inform and shape the judgement in question. As Scanlon notes, the formation of attitudes is "generally constrained by [our] general standing judgements about the adequacy of reasons" (Scanlon 1998: 24).[17]

Since we are responsible for the content of our attitudes in that they constitute our judgements about reasons, we should not take this fact to represent our judgements as 'onetime assessments' (although they can be that). The wide interpretation of RR holds that our evaluative judgements are "not necessarily consciously held propositional beliefs, but rather tendencies to regard certain things as having evaluative significance" (Smith 2005: 251). Such judgements are "continuing and relatively stable dispositions to

[17] For a different but complementary development of this line of thinking see Kumar (1999: 289-292).

respond to particular ways to particular situations" (Smith 2005: 251, n.27). In order to properly identify and fix the content of these dispositions, we will have to place each 'onetime assessment' into the broader "evaluative framework through which we view the world…[comprising] the things we care about or regard as important or significant" (Smith 2005: 251-252).

Given this evaluative framework, Smith argues that "if one judges some thing to be important or significant in some way, this should (rationally) have an influence on one's tendency to notice factors which pertain to the existence, welfare or flourishing of that thing or person. If this is so, then the fact that a person fails to take note of such factors in certain circumstances is at least *some* indication that she does not accept this evaluative judgement" (Smith 2005: 244, my emphasis).[18] As I noted above, with respect to certain other instances of caring or loving, it can even be constitutive of holding an attitude that we exhibit certain extended and varied patterns of awareness or interest in the object of our attitude.[19]

The wide view of RR emphasizes the systematic and relatively stable synchronic and diachronic relations between our various judgements and tendencies in determining what it is that an individual is to be called to account for; the wide view holds that when fixing the content of our judgements for the purposes of moral assessment, we must make reference to our other judgements, tendencies, dispositions and actions. I contend that

---

[18] How much? Answering this question pushes us to examine other elements of the agent's psychological make-up.

[19] While one might claim this is true of any attitude, properly so called, I'm not sure this is the case. There is wide variability in how one might experience and display an emotion such as anger, and it is not obvious to me that we can even say that it is constitutive of 'anger' that if one does not exhibit a certain pattern of awareness then one is not angry. What I want to emphasize at the very least is that there is a tighter connection between certain patterns of awareness and the presence of certain emotions (e.g. caring) than there is for others (e.g. anger).

not only is the wide interpretation not subject to the objections that I detail below, but is a plausible interpretation of the mark of responsible agency.

Section 2.5 – Objections to the Rational Relations View

The first objection holds that RR simpliciter cannot make good sense of the fact of attitudinal conflict, common to cases of practical irrationality, for the purposes of moral appraisal.[20] It is claimed that in situations where an agent is unsettled with respect to her attitudes, the rational relations theorist cannot explain why we ought to privilege one aspect of an agent in determinations of moral responsibility. Akratic action is one example of this challenge – is the agent primarily responsible for her (ignored) judgment or the (weaker) reason she acts upon, or the whole state of affairs? This charge claims that whatever option chosen by RR, we will be arbitrarily privileging one aspect of the person to the detriment of others.

To elaborate: in the akratic case, it seems RR is committed to the thought that while both judgements belong to the agent (that is, her considered judgement and the judgement that runs contrary to whatever the former is), we ought to privilege one of them in our responses to their irrational action (in particular an agent's considered judgement). RR, the thought goes, cannot properly distinguish which attitude, when these attitudes conflict, ought to speak for the agent insofar as both are connected to our evaluative judgement.[21] Given that both judgements nominally speak for the agent, but are conflicting, we have no principled way of determining where the agent lies. Such a situation seems difficult to understand – and while there are characters, e.g. 'Lui' in Diderot's *Rameau's Nephew* who approximates this contradictory existence, exclaiming

[20] For one version of this objection see Shoemaker (2011).
[21] Arpaly (2000)

"The Devil take me if I know at the end what I am. I have a mind as round as a ball and a character as straight as a willow: never false if I have the slightest interest in being true, never true if I have the slightest interest in being false" (Diderot quoted in Williams 2002: 188) – it is hard not to think that we require a perspective that can give an univocal answer to the question, where does the agent stand? The mere fact of being connected to our evaluative judgement is not sufficient to supply a criterion that can order our moral assessments in cases of practical irrationality.

However we answer this preliminary concern, a further worry remains that in true cases of irrationality (if the cases above are merely conflicting rather than contradictory), RR is committed to a rather implausible result. That is, in cases where we judge that, say, a spider is not dangerous while simultaneously feeling afraid of spiders (due to the fact that we judge them to be in some sense a danger), RR seems to be committed to the thought that in order to be held responsible for our irrationality, we must judge that the state of our psychic system (the state of being irrational) is a good thing in itself. In short, if RR holds we are responsible for our attitudes in virtue of our having judged that something is of evaluative significance, then we must judge that the state of holding contradictory, hence irrational, attitudes itself to be of evaluative significance in order to be held responsible for irrationalities.[22] Since we do not seem to make this judgement regarding such higher-order attitudes, it seems RR cannot make sense of how we are responsible for our practical irrationality (Shoemaker 608).

[22] One tack a defender of RR might take in this instance is to grant the objection but hold that since it is not immoral in this instance to exhibit a form of practical irrationality, this is no objection to RR, which is a theory of *moral* responsibility. While I think such a move is possible, I think it unduly limits RR and implausibly claims that all forms of practical irrationality are not immoral.

The second objection claims that the tie between an agent's evaluative judgment and her attitudes and actions cannot be adequate for ascriptions of responsibility because while an agent may possess the general capacities necessary to determine her attitudes, in some particular circumstances in question, she may lack the relevant ability to exercise them. The fact that the general capacities are defeated in the particular circumstances, it is claimed, undermines our intuitions that an agent is responsible, a fact which RR cannot accommodate.

In this vein Neil Levy claims that facts that are morally salient must be 'personally available' to an agent in order for the individual to be held responsible for her attitudes or actions. Personal availability is understood here as the following: "Information is personally available when it is so readily available that it requires little effort to retrieve and it is poised to guide behavior" (Levy 2011: 246-247). So, for example, attitudes which express ill will for which there are good reasons to avoid will not properly express who we are when they are not personally available in the above sense. According to Levy, any expression of character that stems from this attitude would not be caused in the appropriate fashion and so would no longer be able to speak on behalf of the agent.

Why would it lack the required connection? It is because, on Levy's view, it would not be the content of the morally objectionable attitude itself which would cause the morally problematic attitude or action. That is, if I am ignorant of some important moral fact (in that it is not personally available to me), say that lying to one's adult children is paternalistic and wrong, a wrong that may result from my having lied would not properly speaking stem from the above attitude that causes me to perform that action

at the time. It is not the fact that it is wrong, the thought goes, which caused me to commit this objectionable action; rather it is an attitude which exists under a different description that caused me to so act (perhaps that even though they are adults, one believes that they are still not able to deal with the hard truths of the real world).

Now if RR is committed to holding us responsible for our attitudes insofar as they express who we are or insofar as they reflect our evaluative judgement, claiming that we ought to be open to blame or sanctions for this attitude would be inappropriate because it does not actually express our moral agency. It would be unfair, then, to hold someone responsible for what they could not avoid doing due to the lack of awareness of the reasons for action. Fundamentally, this charge holds that a general capacity to recognize and modify our attitudes in response to our judgements concerning reasons is not sufficient to ground a theory of moral responsibility.[23] When a particular ability is lacking, when reasons are not personally available to an agent, something which RR permits, then it licenses an unfair form of responsibility attribution and ought to be rejected for such a reason.

The third charge that trades on the distinction between narrow and wide RR is related to the second in that it holds that the link RR draws between our attitudes and our evaluative judgment cannot make sense of what is often taken to be the most attractive part of the view – how we are responsible for what we notice and neglect, forget and omit. According to RR, omissions, for example, are attributable to us for the purposes of moral responsibility because they reflect our (implicit) evaluative judgements. That is, if I forget to call you on your birthday, this can reflect, in part, the judgement that you do not

[23] It also poses a question regarding the nature of moral knowledge necessary for responsible action as well as raises concerns about fairness. These charges will be considered in later chapters.

matter enough to me to recall such a detail, or to make sure that I do not forget when the times comes around (say by making a note on the calendar). Omissions, according to RR, can be significant insofar as they can convey implicit evaluative judgements of the agent, and can provide good evidence for the attribution of a certain kind of judgement (say, that your birthday is not that important to me). But if the nature of omissions is such that their content is not actually part of our evaluative judgments, it is difficult to see how to properly draw the relevant connection between an agent's judgment and her omitted act.[24] If we take seriously the thought that an omission constitutes an absence, it is hard to see how such an absence can give us any insight into who the agent is and what judgements they do in fact hold.

George Sher writes that it is difficult

> to square the attributionist's [RR theorists] claim that an agent's responsibility is restricted to those features of his acts that reflect his judgments about reasons with the fact that agents often seem responsible for acts whose wrong-making features have not registered with them at all…if the wrong-making feature of what Alessandra does has no input at all into her all-things-considered judgment that she has reason to do it, then it is hard to see how that judgment can possibly connect her to the act's wrongness in a way that renders her responsible for it (Sher 130)

Sher's thought seems to be that we cannot hold an agent responsible for an insensitivity or an ignorant act insofar as their insensitivity or ignorance did not figure in their judgements, which is the mark of responsible agency according to RR. The (wrong) act that they actually did perform was understood under a different description by the agent – what actually figured in the content of her attitudes was not the (wrong) act but something else. As a result, RR, it seems cannot make sense of how it is agents are responsible for what we notice and neglect, forget and omit.

---

[24] For another version of this charge see Fischer & Tognazzini (2009).

Section 2.6 – Responses to the Objections

It should be noted that each of the above charges is plausible; yet I think much of their plausibility stem from a misunderstanding of RR, that is, the fact that we can distinguish a broader and narrower version of the rational relations view. In each case, the charges appear to be plausible if we read the rational relations view as committed to a narrow connection between our judgment and our attitudes and actions, while the broader version does not suffer from these difficulties, or so I claim.

Some philosophers, nominally friendly to the rational relations view, have taken the narrow interpretation to be representative of the best version of RR. George Sher, for one, calls the rational relations view 'minimalist' because it supposedly limits our responsibility to our explicit evaluative judgements. Such a view is minimalist, or narrow in my terminology, due to the fact that it requires responsible agency to exhibit a conscious judgement on the part of the agent which then determines her responsibility or lack thereof. Sher holds that limiting responsibility to these acts of conscious judgement is too minimal to do justice to our actual practice of responsibility attributions. Such considerations push him to argue for a more 'maximalist' version of what can be attributed to individuals involving the psycho-physical structures which underlie our self.[25]

Whatever the merits of Sher's own view, each charge can be rebutted by attending to the distinction between wide and narrow versions of RR.

Recall the first charge – privileging one aspect of the self over another is arbitrary and that RR is committed to the implausible result of attributing an attitude that takes the entire agent's irrational psychic state as having evaluative significance if we are to be

[25] See Sher (2009) *passim*.

responsible for our irrationality. The narrow view does render practical irrationalities such as akratic action confusing, if only because it provides the impression that each conscious and explicit judgement is to be fully owned by the agent. Given that they are inconsistent, the fact that one judgement is 'considered' while the other is viewed as aberrant or to be rejected in a cool hour, *is* arbitrary. This would almost eliminate the sense in which an agent is irrational. Moreover, in attempting to own the entire inconsistent set of judgements, an agent approaches incoherence.

The wide interpretation can handle these concerns. For it can accommodate how both inconsistent judgements, to the extent that they are simultaneously and legitimately held, are attributable to the agent, albeit in differing ways. For presumably insofar as an agent is subject to a momentary weakness, say that smoking one last cigarette won't really violate her commitment to quit smoking, such a judgement on her part will support a smaller and perhaps more insignificant range of dispositions that are otherwise part of the agent's psychology. That is, if the quantity and quality of the connections between a judgement and other elements in her psychological make-up are few, this may suggest that what the agent is responsible for is not revealing some stable and longstanding disposition to return to smoking but rather that she has only momentarily lapsed back into a problematic habit. Looking to the quantity and quality of the connections between a judgement-sensitive attitude and other elements of her psychology, something required by wide RR, can allow for more fine-grained determinations of what exactly an agent is to be considered responsible for. Thus it will appear how we can say that while both judgements speak to who the agent is (the commitment not to smoke at all and the insignificance of having one 'last' cigarette) and what we are to hold them responsible

for, nonetheless one judgement lacks both in significance and what we might term critical mass with respect to 'who they really are'. That is, it allows us to put into perspective whether or not the failing is indicative of or reveals a large or insignificant section of a person's psychology.

In everyday terms, we can make sense of why an instance of weakness of will was either a momentary lapse that does not speak to who the agent is, or alternatively, was a revelatory moment that gave us true insight into the agent's character.[26] The wide view suggests that we take account of other parts of an agent's psychology when we determine just what the content of an agent's judgement is. Far from arbitrarily privileging one aspect of the self over another, the wide interpretation of RR can make sense of our more nuanced and fine-grained responses to irrationality in allowing for both judgements to speak for the individual, but in differing ways.

Further, in terms of our overall responsibility for our irrationality, the wide view need not find itself committed to the problematic idea that we judge the contradictory state of our psychic system to be something worthwhile or having evaluative significance. The wide view holds that our attitudes are fixed by both our evaluative judgement as well as the relations between the other judgements that we in fact or could be seen to (not necessarily consciously) hold. The content of a judgement is not, rationally speaking, independent of the other judgements we are committed to. On the wide view, being attributively responsible for our irrationality involves the possibility of

[26] This does not commit RR to holding that the attitudes that can be attributed to us are, in the first instance, stable dispositions of character. While RR is not precluded from holding that we can (and sometimes are) responsible for states of character, the primary object for RR remains a connection with evaluative judgement, which does not require the presence of a stable character trait in order to be exemplified. For more on this issue, see the discussion of the 'thin' sense of RR below.

coming to an awareness of the inconsistency between the content of our attitudes, something that it demands insofar as we are to be responsible for the attitudes individually. For the contents of the attitudes themselves make reference to other aspects of the individual's psychology. Consequently, we need not assume responsibility for the entire state of our psychic system in order to account for our being responsible for our irrationalities, but only for each of our individual attitudes. Insofar as each individual attitude is part of the broader set of judgements which constitute an agent as a particular concrete individual, when there is a conflict such a situation will push an agent to come to a recognition of it as a conflict between different attitudes that are each separately attributable to her. And it is important to note that the agent will experience a form of 'normative pressure' to resolve this tension with respect to the conflicting attitudes (insofar as they are truly inconsistent). But the responsibility for the set of irrational attitudes will be contained, as it were, within our responsibility for each individual attitude on its own. The wide interpretation, I think, captures and accommodates this fact without having to embrace the implausible judgement outlined above.[27]

Recall that the second charge held that due to the absence of an individual's ability to make certain judgements, in that morally important information is not personally available to an agent, then it would be unfair to hold an agent responsible for the attitude or action in question. The narrow view, again, seems susceptible to the charge, but on the wide view the charge is much less plausible. According to the narrow view, if the content of our judgements is fixed by whatever it is we determine at the time to be sufficient reason to form an attitude or perform an action, then it does seem unfair

[27] This response to Shoemaker's concern follows Smith's own response to his objections. See Smith (2012). For criticism of this type of strategy, see Ross (2012), section 4.

to hold such an individual responsible for whatever objectionable attitude stems from such attitudes or actions. If the view we take of an agent for the purposes of moral assessment is narrow in this way, then the fact that certain reasons are not available to an agent (and so did not factor into their deliberations) even while they retain the general capacity to properly recognize such reasons makes it the case that we should not morally criticize the agent for whatever objectionable act they performed.

I recognize that a defender of RR might simply claim that Levy's charge is misguided in that it confuses two separate and distinct questions. On the one hand, we can ask whether or not an agent is responsible? On the other hand, we can ask that agent should be held responsible? The latter question is separate and distinct from the former. Given that the former is properly speaking the subject matter of discussions of responsible agency, the fact that we may have intuitions about the latter is not immediately relevant to determining who can be a responsible agent. Properly separating the two questions could allow us to see that the background motivation to the 'unfairness' charge is perhaps confused.[28]

Even if the above is true, the concern regarding unfairness remains. Happily, the wide view seems to be less susceptible to Levy's criticism, at least in the first instance. Given the wider range of attitudes which form part of an especially cruel disposition or demeanor, say, it seems both much less unfair to hold such individuals responsible for cruel actions as well as much more plausible that cruel actions are actually caused in the right way by the agent's judgements, even though certain reasons against those actions

[28] Scanlon and Smith offer more detailed elaborations of the distinction as well as a response to the 'unfairness' charge, one that I will take up in more detail in chapter four.

are not available to an agent while they retain their general capacity for reason recognition.

The fact that certain reasons are personally unavailable to an agent can be explained by the critical background mass of attitudes that form an agent. Holding an agent responsible for these attitudes does not seem unfair when cast in this light, arguably, because the unavailability is itself related to aspects of the individual's psychology. Even though they cannot recognize certain reasons, this is not a kind of mental 'block' that is similar to an outside force. but an incapacity that is explained and expressed by their own attitudes.[29] Whether or not this inability is temporary or long-term or leads to typical or uncharacteristic behaviour does not seem central. When we place the personal inability in a broader context, one that stems from and makes reference to other aspects of the individual's personality, as the wider interpretation suggests we do, we gain explanatory resources to make sense of the unavailability. As a result, Levy's concern that it is unfair to consider an agent responsible for an attitude or action that stems from this inability ought to be attenuated. The wide view renders skepticism regarding our responsibility in these instances much less plausible than it may have appeared on the narrow view of RR.[30]

The last charge maintains that if RR is committed to holding that we are responsible for the content of our judgement, because the content of my judgement in the case of an omission or a forgetting is not, say, that I ought to forget or that something else

---

[29] For an elaboration of the nature of these incapacities see Williams (1993).

[30] I recognize that this does not eliminate philosophers' concerns regarding the unfairness of claiming an agent is responsible even though certain information (normative or factual) is unavailable at the time of acting, as this concern has a multitude of sources. However, insofar as this concern is animated by the thought that since, *at the time*, information was not available, taking a broader picture of the agent's psychology that is not limited to a certain time slice should alleviate this source of concern regarding unfairness.

is more important, but that ‘x’ or ‘y’ is the thing to do, then we cannot be responsible for our forgettings or omissions. If I forgot to accompany my child to school or omitted an important note from a document, short of having intentionally brought about these judgements, they do not speak about the agent *qua* person, for the content of the judgement in question by its very nature does not reflect any judgement regarding the reasons for or against forgetting or omitting a particular action.

On the narrow interpretation, such a charge is plausible. For it seems to be the case that it is false that when we judge that ‘x’ is good or best, or bad or wrong, we thereby judge all things ‘not-x’ to be without importance. That is, if we focus on the individual judgement of a cruel person, who takes pleasure in hurting others, the judgement ‘inflicting needless pain is fun!’ does not seem to carry with it the thought that all reasons against the infliction of needless pain lack merit.

The wider interpretation does not suffer from this difficulty – it can accommodate the thought that what makes an individual cruel is the fact their judgement reflects a broad set of tendencies to both dismiss certain considerations as good reasons and to view others (which are in fact bad) as good reasons for action. When, for example, a friend repeatedly forgets important events in your life or a family member often omits to include you in collective activities, the judgements that they do in fact make can (and do) reflect other aspects of their psychology which would seem to imply, at least in part, that you and your interests or feelings are not of importance to them. This wider view which sets our explicit judgements within the larger web of an individual’s broader psychology can make sense of how an omission or a forgetting can in fact factor into the content of a person’s judgement. It can do so insofar as determining the content of an individual’s

insensitive or cruel attitude or action will require reference to other aspects of their psychology and will not be akin to making a 'onetime assessment' of the person.

The wide interpretation then makes good sense of how RR sees our judgements. In this vein T.M. Scanlon has written that "[a] person who is unable to see why the fact that his action would injure me should count against it still holds that this *doesn't* count against it" (Scanlon 1998: 288). Such a claim seems ill-motivated on the narrow view, for it seems to demand both that an agent not explicitly judge 'x' on the one hand, but nonetheless still somehow judge 'x' on the the other. When we consider, however, that our judgements are not isolated capsules that do not come into contact with each other but are, normatively speaking, a set of interconnected tendencies and dispositions integrated into a largely coherent psychology, it becomes more plausible to see the inability here as stemming from the content of the attitude that they do in fact hold.

It is this background which makes sense of the fact that an individual's cruel judgement is only best understood as such in terms of their whole psychological makeup – that is, their entire set of psychological attitudes and judgements (as well as actions, insofar as they are intentional). The wide interpretation makes sense both of how we can fix the content of a forgetting or insensitivity as one that exemplifies the problematic tendencies as well as accounting for our reactions to the actions of particular individuals.

I've claimed that some recent criticisms of RR stem from a mistaken view of what resources it has to offer. In particular, I've argued that the criticisms above target an impoverished version of RR, one that does not recognize the 'wide' role of other aspects of an agent's psychology that play such an integral role in the fixing of the content of a person's evaluative judgement. It is the wide interpretation that is most faithful to the

motivating ideas that animate the rational relations view – the idea that our rational self-governance, for which we can be held accountable, is an integrated set of judgements concerning what we have reason to believe, feel, and do, a set which is an ongoing, extended and relatively stable entity.

Section 2.7 – The Rational Relations View and the 'Whole Person'

RR improves upon other theories of moral responsibility by showing how we can integrate the various different aspects of our personalities, what we might call 'the whole person', within the scope of moral evaluation.[31] Various different philosophers, Harry Frankfurt prime among them, have drawn our attention to the way in which attitudes or actions can be a kind of 'outlaw', that is, present within our psychic system but lacking any and all authority to speak on behalf of the agent. While all accounts of agency will have to determine whether or not some happening in our psychic system is to be counted as part of our agency for the purposes of moral assessment and appraisal, some draw the line in fashions that can be counter-intuitive, excluding worrisome or objectionable attitudes (say, some latent sexist thoughts or a cold indifference to another's feelings) because the individual has wholeheartedly rejected these attitudes as part of their psychic system. Assuming this kind of simple identification or endorsement view of moral responsibility is correct, we can render certain attitudes outlaws by a decisive act of our own, even if these attitudes persist in our personalities and manifest themselves from time to time in our actions. Arguably, this is counter-intuitive; we ought to be receptive to the thought that even elements of one's psychology that an agent does not endorse ought to be considered when we are making determinations of moral appraisal. If this is correct, then RR has an advantage in being able to incorporate different elements of a person's

[31] For elaboration of the idea of the 'whole person' see Arpaly and Schroeder (1999)

psychology insofar as they connected to her capacities for evaluative judgement (whether or not they are ultimately endorsed or endorsable by the agent herself).

This need not mean that the 'whole person' is the object of moral assessment or appraisal every time we evaluate an individual's action or attitude. There are a variety of different concerns that can motivate our evaluative practices – some prudential, some moral, in addition to other factors as well (aesthetic, for example). All reference to the 'whole person' does is to ensure that anything that falls *within* the scope of the whole person is a candidate for moral assessment and evaluation. It also means that those assessments and evaluations will be more fine-grained (and, hopefully, accurate) in bringing in other elements of the individual's psychology.

So while I think the wide version of RR is the best one available to defenders of the rational relations view, there is a related but distinct set of distinctions we can make regarding RR, distinctions that may be suggested by some of my comments above.

In responding to the objection regarding RR's ability to make sense of practical irrationality, I claimed that certain of our judgements possess a kind of critical mass of connections with respect to other elements of a person's psychology, permitting us to either claim that an instance of practical irrationality was a revelatory moment or a momentary lapse. The fact that some of our judgements bear a more obvious relation to other elements of our psychology might lead one to think that in cashing out the nature of the connection that RR posits between our capacities for evaluative judgement and our other judgement-sensitive attitudes, we ought to formulate a test, as it were, where in order for an attitude to count as attributable to the agent for the purposes of moral assessment, it ought to exhibit some minimal level of connection ('critical mass') with

respect to other elements of a person's psychology. Call such a view a thick version of RR.

According to the thick version, the mere fact of connection between an agent's evaluative capacities and her other judgement-sensitive attitudes is insufficient to ground responsibility-attributions. Because an attitude, even though it is rationally speaking related to the agent, displays few if any connections to other elements of a person's psychology, we are less liable, and we ought to be less liable, to view it as speaking for the agent for the purposes of moral appraisal and assessment. Call a version of RR that permits the mere fact of a rational connection between an agent's evaluative capacities and her judgement-sensitive attitudes sufficient to ground responsibility attributions a thin version of RR. While this minimal connection may not in fact hold, given the presence of various types of defeaters such as a blow to the head or some other way in which a person's evaluative capacities are disabled or bypassed, the thin connection itself is enough to ensure that an agent can be considered to be morally responsible for an attitude or action.

The thick interpretation meshes well with an influential line of reasoning concerning the nature of agency that has come to be known as the 'Real Self View'.[32] All 'Real Self Views' (RSV) are committed to the thought that there will be some type of division within an agent. As David Faraci and David Shoemaker note, "contemporary RSV theorists typically incorporate two (or more) levels (e.g., of desires) or motivational systems, such that the real self is located at the level or in the system that can reflect on and govern its will (or 'superficial self')" (Faraci and Shoemaker 2010). While I talk of division and the authors mention levels, the larger point remains – all real-self views will

[32] See Wolf (1990) for the coining of the term and criticism.

make some type of division within the self wherein an agent can either be a candidate for moral assessment or excused from responsible action.

The thick reading of RR provides a plausible place in which to make this type of cut. By emphasizing the number and nature of the connections that an attitude has to other elements of an individual's psychology, it offers an attractive picture of a necessary part of a real-self view.

Indeed this type of cut (thick RR) within an agent's psychology has appealed to writers outside the domain of moral responsibility. Jocelyn Maclure and Charles Taylor, in the context of examining the legitimacy of the grounds of requests for accommodation based on religious beliefs, argue that we can isolate a set of 'core beliefs' that take pride of place in determining who an individual is. Core beliefs and commitments are attitudes that "allow people to structure their moral identity and to exercise their faculty of judgment in a world where potential values and life plans are multiple and often compete with one another" (Maclure and Taylor 76). Maclure and Taylor recognize that not all of an agent's attitudes will qualify for this title. They write:

> core beliefs and commitments, including religious ones, must be distinguished from other personal beliefs and preferences because of the *role* they play in individuals' moral identity. The *more* a belief is linked to an individual's sense of moral integrity, the more it is a condition for his self-respect, and the stronger must be the legal protection it enjoys (Maclure and Taylor 76, my emphases)

While the context of determining the nature and bounds of religious requests for accommodation and the appropriate legal protection it ought to be enjoy is orthogonal to the discussion here, it is notable I think that Maclure and Taylor resort to elaborating a position that is similar to thick RR in order to explain the legitimacy of religious accommodation. Explicitly appealing to the greater role that certain core beliefs play for

individuals (on both qualitative and quantitative axes) they believe they can articulate a sensible division within the self that makes sense of our intuitions regarding moral integrity. Perhaps more importantly for my purposes, they hold that the mere fact that an attitude is rationally connected to our capacities for evaluative judgement (a simple preference for cake over cookies, for example) is insufficient to give this attitude the necessary 'depth' for it to play a significant role in grounding accommodations.

Holly Smith (2011) has explicitly argued for what I am calling thick RR in order to remedy what she determines to be various defects in 'attributionists' position (what I call RR). Smith offers the case of Clara, a woman who dislikes a classmate of hers, specifically the classmate's hairstyle. Clara refrains from acting on this objectionable attitude in order to maintain a proper reputation and to impress her boyfriend with her good character. However, while in her psychology class one day, Clara is hypnotized[33] and her desire to maintain a proper reputation and to impress her boyfriend are 'frozen' as it were, removing their normal role in her psychological and practical life, allowing other elements to come to the fore. As a result, Clara later on cruelly insults her classmate's hair online, posting a mean comment on a photo, embarrassing her in front of her peers.

Smith's intuition regarding this case is that Clara is not responsible for the particularly cruel action she ended up engaging in. According to Smith,

> this [non-responsibility] seems to be precisely because we understand that Clara has other motives that would normally have contributed to her making a decision

[33] While the presence of hypnotism complicates the example to a degree, insofar as hypnotism is often thought to be an excusing condition, the point remains that Smith wants to argue that an agent who acts from such an automatic process is not responsible for her subsequent attitude or action. However, in contrast to RR, it is not the disruption *per se* which poses the problem, but the fact that the disruption prevents the thick self from being expressed. This is what prevents the attribution of responsibility.

> about posting the screed, and that these desires did not play their usual role once she was hypnotized. In other words, although her choice did arise from her own desire, and does reflect her entire evaluational response at that moment to her options, the choice did not arise from anything like a reasonably full configuration of the motives that she actually has and that would normally bear on such a decision. Her choice reflects part of Clara's psychology, but not enough of her psychology to warrant a judgment of blameworthiness (Smith 2011: 134)

Clara's desire did not implicate a sufficient amount of her 'whole person' for it to be intelligible to take her action to be representative of her true self. Lacking such a 'full configuration' entails that an agent would not be responsible for the subsequent attitudes and actions she engages in. The fact that only a single or small number of desires are implicated in the action that Clara actually takes would not provide sufficient grounds, according to Smith, to render a responsibility attribution appropriate. This is because such an attribution would not shed adequate light on who the agent is, or their broader psychological makeup, for the purposes of moral assessment.

There seems to be some motivation for moving towards a thick version of RR. Smith's view that a choice must reflect a large or full proportion of an agent's psychology in order to hold her responsible for her choice, in particular, seems to offer strong reason to favor the thick interpretation of RR over its thinner rival. That said, however, I think that the thinner interpretation of RR is both more plausible as a version of RR and a better fit for our practice of holding agents responsible. Reflecting on such practices, I think, will show why the thinner interpretation is preferable to the thicker.

Section 2.8 – The Rational Relations View Through Thick and Thin

Prior to examining these practices, I want to briefly note that favoring a thinner interpretation of RR is arguably more faithful to the basic motivations of the theory. The thin interpretation emphasizes that there is a rational connection between an agent and

her attitudes (even if the extent and quality of the connection is not as robust when compared to others) and that this is what renders an agent open to moral appraisal and assessment.

The thought that if an attitude does not reflect or properly express an agent's whole personality then it should not be considered in a moral assessment, in my view, seems to ignore the fact that attitudes that do not reflect large swaths of our psychology still do bear a morally relevant and recognizable connection to our evaluative capacities. They do seem to indicate some thing, however small or insignificant, about a person. While this level of significance can be incorporated into the type of appropriate response to the agent, it does not seem to obviate the fact that the agent can be responsible for the attitude itself.

Turning now to our practice of holding agents responsible, reflection on such practice, I think, offers support for a thinner version of RR. For example, consider an agent, Alan, who after a period of doubt and anxiety, undergoes a change in perspective and subsequently performs an uncharacteristic and unusual action, say abandoning his prior political convictions and turning 180 degrees opposite in his new political orientation. While such radical alterations in attitude and action are not common, they are also not impossible or unknown. While the new judgement espoused by an individual, say "The Old Party, of which I was formerly a devoted and fervent adherent, is rotten to the core" bears few, if any, relations to the vast majority of the other judgements that an individual otherwise might possess, it does not seem to be such an outlier that would render it ineligible for attribution for the purposes of moral responsibility. Nor do we need to make reference to those judgements in a person's psychology in determining the

content of the new judgement, just because it bears so few connections. However, if the new judgement was formed by a person's rational capacities, even though it does not reflect what we might otherwise take to be a person's overall, all things considered or cool and reflective account of what might be true or good or right, nonetheless it seems that such a new judgement should still be a candidate for being an object of moral assessment. The thicker interpretation, then, cannot make sense of our practice of responsibility attributions in cases of radical character change.

Of course such situations are confusing and are certainly coloured by how individuals see themselves as well as by the larger narratives that we often sketch to make sense of and order our lives. These activities though presuppose that the attitude in question is not an outlaw in the sense of being some kind of foreign force that has taken over the individual in question. Such activities seem to demand that the agent actually be open to discussion and criticism, and be able, at least in principle, to have requests for reasons directed to why they have repudiated their previous views and taken on a new one.

This is a weaker level of integration than what might be otherwise demanded. It allows for various types of anomalies and aberrations with respect to individuals we think that we know very well. Indeed, it allows for the possibility that individuals surprise themselves when faced with a situation that is unfamiliar and difficult to anticipate. While a person may have structured their life around one set of responses to some circumstance, considering say an unexpected moral dilemma or having to decide how best to manage or prioritize a work/life balance, when 'the chips are down' we might ultimately side with the option that we had previously thought to be unattractive (or,

even, not personally possible or volitionally impossible, if there is such a type of possibility).[34] Responsibility can still take hold here, and agency has a grip, both from our own perspective and from the perspective of others. We should be forced, I think, to admit to ourselves that 'x' is part of our self even if it is insignificant in the quantity of beliefs it is connected to or the significance of the role that it might play. A thin interpretation of the responsibility requirement of the rational relations view accommodates such facts.

More fanciful or science fiction-style cases pose other problems. Scanlon considers the case of a mad scientist who 'injects' a belief into the psyche of an individual, bypassing the individual's rational capacities, by stimulating their brain such that they form a deep-seated hatred for certain people and see their being harmed as something to promote. Scanlon does not think that, at least in the first instance, that the presence of such feelings should be seen as morally significant, or as candidates for responsibility attributions.[35] One reason Scanlon holds this is that the presence of such feelings "does not tell us anything interesting about this person. Anyone would react in the same way" (Scanlon 2002: 174). The fact that such stimulation would impact any individual in a like manner says that there is nothing unique to the attitude as an actual reflection of the particular individual's personality. In effect, there is no need to resort to other attitudes of the individual to fix the content of the belief, because the individual

[34] For the idea that 'chips are down' situations are especially revelatory, see Frankfurt (1988a: 85).

[35] If we interpret H. Smith's example as being similar insofar as hypnotism also bypasses an agent's rational capacities, then the distance between the thin interpretation and the thick interpretation may not be as great as I've made it out to be. That said, what seems to be driving H. Smith's intuitions, as I argued above, was not the fact that the person's rational capacities are bypassed but rather that her present psychological states are bypassed, regardless of the connection of the new attitude to our rational capacities. Thus there remains a difference in the reason that Scanlon and H. Smith both see hypnotism as an excusing condition for an agent.

could have any complex of attitudes whatsoever and the content of the belief would still be the same. The expression of such a belief would tell us nothing about a person's judgements about reasons (or their more general standing judgements that structure our more particular pronouncements). Nonetheless, should such an attitude persist over a fair bit of time and survive within an individual's psychology, Scanlon holds that it would be a legitimate candidate for attribution and assessment. This is because it would be the kind of state that is, in principle, open to requests for defense or modification on the basis of reasons. The fact that the state would possess and begin to potentially manifest this content over time would allow for it to be a legitimate candidate for moral assessment just because it would begin to develop the minimal sorts of connections to our other attitudes that the thin version recommends. We would begin to be able to see the attitude as an expression of the individual person, just because of these connections to an individual's rational capacities. While this is admittedly a limiting case, it shows how the thin version of RR would accommodate the type of hypnosis case that Smith offered above.

In conclusion, I have argued that the best version of the rational relations view holds that we take a wide perspective with respect to filling out the content of our attitudes and that our responsibility attributions ought to be thinly dependant on the quantity and quality of the connections that are present with respect to those contents. Insofar as these judgements are connected to other aspects of our personality, we need to keep these other elements in view when determining just what it is that a judgement incorporates. However, even if we are unable to find robust and stable connections to other elements of a person's psychology, insofar as there remains a thin or weak

connection between a person's rational capacities and their subsequent judgement (even if not conscious), this qualifies such a judgement as open to requests for justification and hence open to responsibility attributions.

I hope to have explained and adequately motivated the RR view, distinguished a number of different possible versions of RR, rebutted a number of possible objections and misinterpretations of the view and offered reasons to prefer a wide and thin version of RR rather than a narrow and thick version. In the next chapter, I examine what I take to be the most serious competitor to RR, Harry Frankfurt's account of responsible agency. While accepting RA, he offers a radically different account of responsible agency. To this alternative, I now turn.

Chapter Three: Frankfurt's Alternative Account of Agency and Responsibility

Section 3.1 – 'Volitionism' and Responsibility

In this chapter I want to set out and criticize Harry Frankfurt's account of responsible agency. He has formulated a view that I will call 'volitionism'. It places acts of agency at the heart of determining when an agent is to be identified with a particular attitude for the purposes of moral assessment and appraisal. Indeed, one writer describes Frankfurt's views as having such "considerable intuitive appeal—so much so that it has, I think, become tempting to think of identification with the motives on which one acts as a sufficient condition for autonomy" (Westlund 484).[36]

In the first part of the chapter I will set out Frankfurt's views in some detail, after which I will criticize his most recent formulation of the distinctive kind of agency that is required for moral responsibility. Along the way I will try and show how wide and thin RR (or just RR from here on in) makes better sense of cases and can provide a plausible account of the structure of an agent's practical reasoning. I close the chapter by diagnosing the difficulties with Frankfurt's account as stemming from an initial theoretical choice to assume, as it were, that the question as to where the agent lies must be a question that the agent poses to herself prior to there being any sense as to who an agent is for the purposes of moral assessment. Frankfurt, I think, does not sufficiently take to heart in what sense we are active and in what sense passive with respect to various aspects of our lives.

---

[36] While Westlund's claim is framed in the language of 'autonomy' I think there is much conceptual overlap between debates regarding the conditions of autonomy and the conditions necessary for moral responsibility. Even if this is wrong, I think the substance of her claim can apply to Frankfurt's influence on the moral responsibility literature.

One preliminary point regarding the taxonomy of views of responsibility should be examined prior to examining the details of Frankfurt's account.[37] 'Volitionism' has

[37] Before examining Frankfurt's account in detail, one concern regarding the interpretation of his work must be dealt with straightaway. In a commentary on Frankfurt's work, Scanlon has written that Frankfurt's more recent writings have tended to focus on "an ideal of psychic health" to the exclusion of questions concerning "an agent's 'ownership' of his or her desires as a precondition of moral appraisal" (Scanlon 2002:167). In effect, Scanlon is claiming that while a focus on the preconditions of agency required for moral responsibility characterized Frankfurt's earlier work, this concern has moved into the background and focus has shifted for Frankfurt to the value of certain attitudes – say, caring or wholeheartedness - one can take towards elements of one's psychic life.

While there has been an indubitable shift in the focus and emphasis in Frankfurt's more recent writings towards discussions of wholeheartedness from his earlier work, a shift I will comment on below, this concern is still governed by the question as to when an attitude can be properly said to 'belong' to an agent. This 'belonging' question is, nonetheless, a precondition for any kind of moral appraisal or assessment that we might otherwise want to make of an agent. This can be seen, if only negatively, in Frankfurt's contentions that various outlaw attitudes, attitudes that have been decisively rejected by the self, are not able to speak authoritatively for the agent. Given that these attitudes cannot speak for the agent, it would be inappropriate to hold her responsible for whatever flows from those attitudes (or to respond to the attitudes themselves). As such, it seems to me that Frankfurt's work implicitly endorses RA, as set out above, and still holds that the conditions for the attribution of an attitude provide the conceptual framework in which we are to answer questions of responsibility.

Whether or not the above is correct, the judgement that Frankfurt's work has been an extended attempt to elucidate just what it is that makes us responsible agents (and thus is concerned with moral responsibility) is confirmed by some of Frankfurt's own writings. In "Alternate Possibilities and Moral Responsibility" (Frankfurt 1988a) Frankfurt considers and famously rejects what he calls the 'Principle of Alternate Possibilities', that is, the claim that being able to do otherwise is required for attributions of responsibility. In the background of this discussion is the thought that what is required for responsible action is personhood, understood in terms of identification with second-order desires. Given the close conceptual ties between the two accounts (personhood and moral responsibility) it seems reasonable to say that regardless of whatever else Frankfurt may have had in mind in formulating his project, a concern with the preconditions of moral responsibility have always been present.

In this vein, John Martin Fischer notes: "It is thus indisputable that Frankfurt considers his account of acting freely to be helpful in making progress in traditional debates about causal determinism, freedom of the will, and moral responsibility. Whatever else Frankfurt had in mind in putting forward his 'hierarchical' account of acting freely, he certainly sought to make contributions to the traditional metaphysical debates about determinism, free will, and moral responsibility" (Fischer 2012: 167).

What can account for Scanlon's doubts is the fact that Frankfurt has become dissatisfied with what he calls the "excessively pan-moralistic approach that many philosophers take to issues concerning practical normativity…What morality has to say concerning how to live and what to do is important, but its importance is often exaggerated" (Frankfurt 1999a: X). Broadening the scope of his investigations to include more personal and religious ideals has pushed Frankfurt to focus less on what we might call 'narrowly' moral questions, and to focus more on the role that personal ideals and projects play in determining our agency. I leave it as an open question as to

become a handy label for a family of different views that seek to account for responsible agency in terms of some type of actual or possible volitional act on the part of the agent. Given the variety of different ways that we can perform volitional acts, it should be no surprise that it is possible for there to be disagreements amongst volitionists as to the best account of responsible agency.[38] While I won't engage in a large-scale discussion, canvassing the various strengths and weaknesses of volitional accounts, they are all united by a concern that our agency must fall under our 'voluntary control' in order to qualify as a candidate for moral assessment and appraisal.[39] For example, if John chooses to eat a sandwich, given that his choice is under his control, and there is no compulsion or deception which might undermine his exercise of his control, then John's choice and subsequent action qualify as candidates for moral assessment.

Those attitudes and actions that are in some sense impervious to our voluntary control – non-voluntary attitudes – are to be rejected as being able to speak for the agent. In this sense, RR is an account that incorporates non-voluntary elements of our psyche as legitimate candidates for moral appraisal and assessment. While Frankfurt's more recent writings have emphasized how we can still be identified with the more passive parts of our psychology, I think it is legitimate to view his account as volitional in this broad

---

whether or not Frankfurt is correct in his complaint regarding the contemporary character of discussions of practical normativity. However we answer that question, I think we can still fruitfully consider his work as an attempt to work out the conditions under which an agent can be held morally responsible for her attitudes and actions, even if Frankfurt would also want to consider various other kinds of responsibility along the way.

[38] Levy (2011) for example differs from Bratman (1999), who differs from Fischer and Ravizza (1998), all of whom differ from libertarian accounts of free will.

[39] See Smith (2005) for a more extensive treatment of the different kinds of volitionism. Michael McKenna captures the central idea behind volitionism as a thesis regarding what is voluntary, and so under an agent's control and thus what is the appropriate target of moral responsibility. He sums up this family of views in a thesis he calls 'Voluntarism': "The only objects of direct moral responsibility are free actions, where free entails all that is required for the control condition for moral responsibility" (McKenna 2008: 30).

sense, as I will argue below. The basis of the disagreement between RR and volitional accounts turns on the role of distinctive aspects of agency must play as a precondition for moral assessment and appraisal. In other words: is choice necessary for moral assessment and appraisal, or can another sort of agential activity fulfill this role?

Section 3.2 – Frankfurt's Hierarchical Account of the Self

Frankfurt's early attempts to isolate the essential features of agency turned on the idea of 'second-order desires'. First-order desires can be understood as brute urges that agents simply possess in virtue of experiencing them. They are the psychic 'raw material' that an agent initially discovers in her self.

Second-order desires, by contrast, take a first-order desire as their object. That is, if a person wants to want something, then she possesses a desire of the second-order. The reflexive and hierarchical aspects of this mental attitude are noteworthy. For second-order desires are essentially reflexive in that they take another attitude, a first-order desire, as their object. The agent who has this attitude must be able to reflect, in some sense, on the contents of her own mental life, before being able to form a second-order desire. And they are hierarchical insofar as we can conceive of second-order desires as articulating a particular structural relationship between our attitudes. In adopting a second-order desire, we stand back, as it were, and assume a certain stance toward the contents of our mental life (e.g. first-order desires). In standing back, or reflexively considering our first-order desires, the agent makes a determination as to the "desirability of his desires themselves" (Frankfurt 1988b: 17). This notion of a hierarchy is also suggestive in that it points to the fact that an agent who assumes a reflective, hierarchical stance towards herself is one in whom there is invested a certain amount of autonomy or

authority. This agent, one might think, can speak authoritatively in regards to which of her desires she wants to push her around.[40] For Frankfurt, "one essential difference between persons and other creatures is to be found in the structure of the person's will…[for] it seems peculiarly characteristic of humans, that they are able to form what I shall call 'second-order desires'" (Frankfurt 1988b: 12).[41] The key to determining both how an agent can be constructed out of the raw material that we find ourselves beset with, and what an individual person amounts to at the end of this construction process, is to be found in the structural features of an agent's will. For Frankfurt, agency and personhood are constituted by the boundaries of an individual person's will and thus map the territory in which we can make attributions of moral responsibility.

It is only in moving to desires of the second-order that can we identify what the person really wants, Frankfurt holds, because it is only then can an agent determine which first-order desire she wants to move her to action. We might say: in so far as an agent acts, her action must flow from her 2nd order desires. Action that flows from first-order desires is a mere driving force that does not express the personhood of the agent – it is not free action. In taking an interest in which desire wins out, which desire she wishes to move her to action, the agent identifies herself with one of the competing desires; the other competing desires, if they remain, become outlaws or alien with respect to the

---

[40] As Frankfurt writes elsewhere: "Our most elementary desires come to us as urges or impulses; we are moved by them, but they do not as such affect our thinking at all. They are merely psychic raw material…Impulses and urges have power, but in themselves they have no authority. They move us more or less strongly, but they make no claims on us" (Frankfurt 2002:184).

[41] In a situation where an agent has two conflicting first-order desires, say a desire to have an alcoholic drink and a desire to not have an alcoholic drink, there is a question as to where the agent actually or ultimately 'stands' with respect to the question: would you like an alcoholic drink? The agent who endorses one of her first-order desires via a second-order desire has resolved this tension and in this sense expresses where she stands on the issue. In this sense, the desire can now speak with authority with regard to the practical question as to what she will (attempt to) do.

agent. While they may not disappear, as they are not endorsed by the agent, and are in conflict with the attitude that the agent does endorse and identify with, they can be seen not as reflections of her personhood but as alienating or external forces in the agent's mental life. Those attitudes with which the agent identifies become internal to her will for the agent has made a decision with respect to how her 'will' ought to be shaped. And when an agent is successful in being moved to action by her second-order desire, this constitutes that agent as a person in Frankfurt's sense. While before the agent was a wanton (Frankfurt 1988b: 16), understood as a being who is indifferent to what moves her to action, in identifying with a second-order desire and wanting that desire to be effective in action, an agent becomes a person, one who is concerned or interested in the course and shape of her will and her life.[42]

Frankfurt's example of the two addicts can help illustrate this view. Consider two drug addicts, one who willingly accedes to her addiction while the other does not. The willing addict is, on Frankfurt's account, a wanton, because she takes no interest in whether the first-order desires that move her to action are those that she wants to move her. Her lack of interest in the structure of her will renders her unable to be understood as a person, because there is nothing with which to identify her. There is no sense in which any of the attitudes we might find in her can be said to be internal to her agency.[43]

---

[42] Thus for Frankfurt, strictly speaking, it is not the mere presence of a second-order desire that constitutes a person. It seems possible, if somewhat bizarre, for an individual, A, to only have second-order desires regarding the mental life of another individual, B. Rather, it is the presence of this desire whose content determines which first-order desire will authoritatively speak for the agent herself that constitutes a person (regardless of whether or not the first-order desire is actually realized in action).

[43] 'Internal' in the sense that via her second-order desire, she has made the first-order desire 'her own'. The contrast between internal and external is not given by the skin or the skull. It refers instead to the process of the constitution of the person whereby, in forming second-order desires,

The unwilling addict by contrast is interested in which desire moves her to action. Beset by the first-order conflict of desires both to seek out and to not seek out a particular drug, the unwilling addict takes a side as to which desire should move her to action. By identifying herself with one desire and forming a volition in order for it to be effective, the agent takes an interest in her will. This activity, when successful, is what constitutes the structure of her will. Rather than expressing the passive repose of the willing addict, the unwilling addict is active insofar as she takes an interest in what her will should look like. She rejects the conflicting desires that are contrary to her second-order volition, rendering them external to her will (Frankfurt 1988c: 67). The activity of identification and internalization constitute the structure of the will, determining the essential characteristics of an agent's personhood. "It is these acts of ordering and of rejection – integration and separation – that create a self out of the raw materials of inner life" (Frankfurt 1988e: 170).

How does an agent identify with certain attitudes, thereby forming second-order volitions? Frankfurt's early work took this process to be a kind of decision. Noting that the etymological meaning of decision is 'to cut off', Frankfurt held that when we decide a desire is to be our own, we make up our minds by cutting off any further deliberation, hesitancy or uncertainty. We become wholehearted with respect to the attitude in question, siding with the one participant in the conflict over another. A conflict at the 2nd order level can exhibit this process. If I am trying to determine which feeling I have towards my friend, and I experience a conflict, feeling both resentful and grateful at the same time, Frankfurt holds that we can resolve such a conflict by making up our minds.

---

an agent identifies with (or renders internal) some aspects of her mental life (e.g. first-order desires) and withdraws herself from others (rendering them external).

This mental act is a decision on our part which resolves our will. We thus identify with whichever attitude we have decided upon and eliminate any uncertainty with respect to it – we are wholehearted in relation to the attitude in question. The other participant in the conflict may not disappear – my resentment, for example, could persist – but the attitude would no longer be mine in the robust sense that Frankfurt thinks is necessary for personhood. The attitude of resentment then would not be properly attributable to the person in question, for she has rendered such an attitude external to her will.

Incidentally, we might note that this account of wholehearted commitment as constituted by a decision on the part of the agent is meant also to explain the authority such commitments have for an individual. We might ask – why is this commitment or desire to be identified with the agent? Why not another? Frankfurt responds: because the decision literally cuts off further discussion or debate, 'resounding' through their mental life (Frankfurt 1988b: 21). When this occurs, the decision to identify is synonymous with our constituting ourselves as that kind of person. This act of constitution at once determines the shape of a person's identity (*qua* responsible agent) as well as grounds the authority of the desire or commitment for an agent – they provide her with authoritative reasons for action.[44] Such an agent could no longer find grounds for reasons of

[44] Frankfurt seems to accept a form of internalism regarding reasons for action. While I will discuss internalism in greater detail later in the work, as a first pass, this is the thought that in order for a consideration to count as a reason for action it must be connected, in some sense, to the contents of an agent's psychological life. For Frankfurt, whenever we are wholeheartedly identified with a care or concern, we, *as a result* of the identification with this care or concern, have reasons to care and be concerned for the objects of our interest. Thus if I identify with a desire to play baseball, I now have reasons to practice baseball, to worry about other people's opinions about my play and the game itself, reasons someone who does not share my desire may not possess. While this, on its own, does not show externalism to be false, I think Frankfurt thinks that it shifts the burden of proof away from the internalist in the dialectic.

resentment towards her friend, because she would recognize those considerations as external to what her actual attitude toward her friend is.

Our commitment, then, is to the attitude, and when we are wholehearted with respect to it, the continuing existence of conflict is no threat to the agent's will, because now it is a question of a conflict between the person and some external force, rather than a conflict internal to the will. Further, this commitment on our part makes sense of the idea that we are responsible for our attitudes – that is, we must take responsibility for our self in order to qualify as a person.

This is Frankfurt's early view. Note however the absence of any mention of evaluation or evaluative judgment in his account. The role of reason, as well as the emotions or feelings, is subordinate to the role that the will plays in the constitution of the person. The formation of our 2nd-order attitudes need not follow the dictates of reason or our emotional natures. While our rational and emotional faculties may be necessary features of personhood, they are not sufficient. It is only in virtue of a structured will that an agent can be said to be a person.

At times, however, Frankfurt speaks as if second-order volitions are evaluative attitudes on the part of the agent, reflecting her reflexive attitude towards the various aspects of her mental life. 'Evaluative' refers to an agent's determination of what is good. So if my second-order volition was to have a particular first-order desire move me to action, an impartial observer could see me as expressing the evaluation that I hold the first-order desire to be good. Second-order volitions could then be viewed as expressions of her considered evaluative preferences for her will – at the very least, evaluation and

judgment would play a critical role in the formation of a person and the grounding of her reasons for action.

Section 3.3 – Frankfurt's Account of the Caring Self

The possibility that evaluation could play a role in the constitution of agency marks the point at which Frankfurt's later writing comes into the picture. For he is concerned to avoid any association between the constitution of our agency and the thought that who we are necessarily reflects our view of the good life or what is best, most appropriate or fitting. This concern to avoid evaluation – especially moralization – in discussions of the grounds of responsible agency are prominent in his more recent writings.

Frankfurt's later work begins by expanding on the basic idea of desire used above. Frankfurt claims that we can analyze what he calls different 'modes' of desire, specifically, what we care about and what is important to us, as well as what we love. Our agency lies in the importance of what we care about. As a result, Frankfurt turns to analyzing the notion of caring as the foundational commitment of responsible agency. Now while Frankfurt has taken to analyzing loving as a particularly important instance of caring, I will not comment upon the distinction, since I don't believe it important for my purposes.

If we focus on the concept of endorsement, we can shed light on Frankfurt's attitude towards evaluation. Endorsement has often been understood as an evaluative concept. That is, if we are taken to be endorsing something, then we seem to be committed to the thought that we have some kind of positive evaluative attitude towards the object of our endorsement. Thus, if I endorse my desire to eat chocolate cake, we

could understand my endorsement as expressing my positive attitude towards this tendency of mine, or a preference of mine towards this particular fondness. Frankfurt, at least in his early work, seemed to favor a thin but recognizable view of the nature of what we endorse: "Second-order volitions express evaluations only in the sense that they are preferences" (Frankfurt 1988b: 19 n.6) he notes. That is, the volitions with which a person was identified could be seen as expressing a preference for the object of the attitude.

And yet while the concept of endorsement seems to carry with it the idea that an agent has some kind of positive attitude towards the object of her endorsement, this does not form part of what Frankfurt has in mind by the term in his later work. He holds that one can endorse something without thereby being inclined to positively or negatively view the object of the attitude. Rather, for Frankfurt, endorsement primarily consists in a kind of neutral accepting attitude toward the object or mental state in question. Acceptance, here, is related to the idea of satisfaction. If we are satisfied with an attitude, then we can be said to accept it, for we can see no further reason to question the attitude. Thus it is acceptance, and not judgment or evaluation, that is key to understanding endorsement – if an agent 'accepts' an attitude, then she can be said to endorse it as part of her person (Frankfurt 1999b: 113).

But such an acceptance need not entail any positive attitude towards the object, and could perhaps coexist with a condemnatory attitude towards the attitude in question. Indeed, acceptance or satisfaction consists primarily in an absence. That is, an absence that indicates there is no desire to alter the attitude in question on the part of the agent. This absence is not gratuitous however. Frankfurt holds that satisfaction stems from a

certain kind of understanding and evaluation of an agent's own psychic situation: "the fact that the person is not moved to change must derive from his understanding and evaluation of how things are with him" (Frankfurt: 1999b: 105). This connection between our satisfaction and our evaluative capacities is required so that the satisfaction in question does not come about through some accident (that is, say, through inattention or forgetfulness) or through the intervention of some kind of force, such as the onset of severe depression. Were the absence of an interest to change to arise in these fashions, Frankfurt, it seems, would not want to say that the agent should be identified with them, even though she would be satisfied with the attitudes in question. So as long as the satisfaction is somehow derived from our evaluational and cognitive capacities and our understanding of what is present within our psychological system, then we can say that an agent is to be identified with the attitude with which they are satisfied. It is being satisfied with what we find within our psychologies that constitutes our responsible agency (and thus is consistent, I believe, with RA).

What we are satisfied with assumes a foundational role in the constitution of our agency and practical reasoning. They are our volitional necessities. Frankfurt argues that "[o]ur essential natures as individuals are constituted, [accordingly], by what we cannot help caring about. The necessities of love, and their relative order and intensity, define our volitional boundaries. They mark our volitional limits, and thus delineate our shapes as persons" (Frankfurt: 1999d: 138). These volitional necessities constrain our choice – they involve our inability to choose or refrain from choosing particular things. They form the boundaries of our will.

These necessities are both contingent, in that they can change, and particular, in that they are peculiar to each individual person. While Frankfurt thinks we have essential natures as individuals that are indicated by our volitional necessities they can and do vary from individual to individual. Thus, they are personal in a strong sense of the term.

These volitional structures also exemplify a kind of psychological necessity – we experience them as a passive force that determines what we are unable to bring ourselves to do. But they are not to be seen as compulsive in nature – for, according to Frankfurt, they stem from our very own will. And insofar as we are wholehearted with respect to what it is we care about, our volitional necessities will form a coherent and integrated structure of the self. They define who we are for the purposes of assessment and appraisal. While these necessities can alter for an individual – they are not permanently fixed – they can only change due to external circumstances or due to more indirect, deliberate efforts. But if the necessity is such that it is unthinkable to alter the necessity, then it will not even be possible to contemplate indirect methods of change. Only external circumstances could give rise to a change in our identity (Frankfurt 1999c: 112).

These volitional necessities also function as the basis of our practical reasoning. In determining how we ought to live, an individual is faced with a variety of tasks. They must decide what they want, relate those wants to their other preferences and they must also determine what it is they value. But individuals also face a distinct and more basic task: individuals must determine what it is they care about. The primacy of this task stems from the fact that without some determinate set of cares, an individual will lose the bearings necessary for effective choice. Without the context of choice constituted by our particular cares, an individual faces a kind of vertiginous situation – a freedom where not

only can we deliberate about to do, but who to be as well. In principle, a limitless freedom provides an agent with the possibility of radically re-designing her situation as well as herself, altering her personal characteristics. Without these individual characteristics, though, we have no principle of choice that can authoritatively guide us in our practical deliberations, for we have no basis upon which to make our decision.

Section 3.4 – Some Concerns for Frankfurt's Views

Before continuing, I want to note some preliminary difficulties that may be thought to be present for Frankfurt's account, insofar as it incorporates elements of volitionism. As Angela Smith has ably demonstrated, there are a wide variety of our actual practices of responsibility attributions that resist being understood, at least initially, in volitionist terms. She notes that we standardly attribute a large degree of significance to what might be termed 'spontaneous attitudes'. Smith writes:

> We care…not only about what people intentionally choose to do, but also about the general attitudes they have and express towards us. Indeed, we often respond to people's spontaneous attitudes, reactions, and unreflective patterns of awareness in many of the same ways that we respond to their voluntary actions (Smith 2005: 241)

This is because, as I argued in chapter one, these spontaneous attitudes, reactions and patterns of awareness are part of 'the whole person', that is, they are properly connected to our capacities for evaluative judgement such that they open up an agent in principle to be held responsible for this class of attitudes.

Smith canvasses a number of different kinds of spontaneous attitudes (or involuntary responses) in order to argue that RR has more theoretical resources to account for our actual practice of moral responsibility than a volitionist. So, she notes, what agents notice and neglect does not standardly seem to be something over which we

have any direct control, nor is it something that we standardly willingly engage in. These are paradigm passive activities on the part of the agent. Yet, Smith argues, "we often take what a person notices and neglects to have an enormous amount of expressive significance" (Smith 2005: 242). The fact that an agent does not notice someone's discomfort or that a comment or action is unwelcome is legitimately understood as 'involuntary' insofar as we do not engage in these activities under this description. Nonetheless, they are standardly thought of as legitimate bases for moral (and other kinds of) criticism, whether or not they are endorsed, or stem (or connected to) some more basic feature of the self, such as a volitional necessity. The reason that they are so thought, according to RR, just is the fact that there is a rational connection between what an agent notices and neglects and other aspects of her psychological makeup. Given the presence of this connection, we can make a legitimate inference as to what an agent cares about, at least in part, based upon what it is that she notices and neglects.

Another class of cases that Smith identifies are those where certain thoughts do or do not occur to an agent. Discussing a case made famous by Bernard Williams, where a businessman brings up and immediately dismisses the far-fetched possibility of murdering a competitor in order to facilitate his business operations, Smith notes that what is disturbing is the fact that the agent in question does not seem to wholeheartedly endorse the thought that the killing of others (except under certain kinds of specified circumstances) is impermissible. While the agent may decisively reject this plan of action as not being one we ought to ultimately follow, the fact that it was even entertained suggests that the agent's commitment to the judgement that unjustified killing is wrong is either weaker than we might have first thought or qualified in problematic and potentially

objectionable ways. Preferably, we might think, this type of thought ought to have never even occurred to the agent (Smith 2005). That what occurs to someone is evidence for what is the actual content of the judgements they possess again is vindicated by the thought that there is a rational connection between our judgements and what otherwise strikes them as a reason for action (even a bad one), what Scanlon elsewhere calls a "seeming" (Scanlon 2002: 178). What strikes as potential reasons for action can give us insight into the contents of other of the agent's attitudes and is certainly relevant, one might think, for the purposes of appraisal and assessment. To illustrate, Scanlon discusses the case of a teacher deciding who to cast in the school play. One of the students' fathers is a hated rival of the teacher and the teacher would not like the pleasure that the father would feel from the fact that their child has been cast in the play. Coming to their senses, Scanlon writes:

> I [the teacher] may judge that this is not in fact a good reason to deny the child the part. I may feel only "disapproval" of the "motivational tendency" of this vengeful thought, and no desire to be moved by it. Yet it is crucial to the "motivational tendency" that it retains that when I think of the play, the pleasure the father would derive from seeing his daughter in the limelight, keeps presenting itself to me as a reason to prevent this from happening. The claim that this desire has on me is not a matter of my approval or endorsement, but of the fact that it consists in something seeming to me to be a reason, even though I judge that it is not (Scanlon 2002: 178-179).

Finally, Smith outlines some further instances of non-voluntary reactions that seem to possess important expressive significance. Forgettings and omissions are some such instances. When we forget an important event, such as a friend's birthday, the regret and shame that an agent might feel for having forgotten can be explained by the fact that we tend to think that if a certain event is important to us, we will not be too quick to let it slip from our radar screen so to speak. And the fact that we attempt to make amends and

apologize for these lapses also supports the idea that we view such instances as ones that count, as it were, as reflections of our moral character. Involuntary responses are other instances of non-voluntary reactions that seem to possess expressive significance. Jealous reactions on an agent's part in response to the deserved success of another may not be chosen under that description but they are revelatory of an agent's underlying commitments. The contents of her attitudes can be better fixed and understood, by an agent as well as others, by how these judgements manifest themselves in our involuntary reactions to different circumstances and events. Again, that these involuntary reactions provide a basis for moral praise or blame seems undeniable – whether we want to praise or blame Meursault in Camus' *The Stranger* for not exhibiting any emotion at the death of his mother, the fact that he is so passive in the face of such a dramatic event seems to speak clearly for many about who this individual is. Indeed, the other characters in the novel take it as good evidence for Meursault being a seedy and unfeeling character (whether or not we agree with this judgement is beside the point). And this fact is best explained by the idea that there is a close and meaningful rational connection between the various ways we manifest our attitudes and actions, even involuntarily, and our capacities for evaluative judgement.

These are just some points that favor a non-voluntary conception of moral responsibility over its volitionist rivals, showing how it can account for a broad swath of our actual practice of holding agents accountable through the use of its central theoretical resources. It should also be noted that RR can straightforwardly account for more conscious acts of agency as well without requiring any special modification of its account of responsible agency.

Given these points, is the view that Frankfurt sets out above sufficient to dislodge RR from being our preferred account of responsible agency? While Frankfurt does not argue directly for his picture, its suggestiveness and intuitive resonance, along with its extended detail, force us to either accept his view or articulate an alternative. This is what I propose to do by first criticising Frankfurt's view and then sketching how RR can replace his account. In the remainder of the chapter I will attempt to substantiate this claim.

Frankfurt's account has been subject to a number of criticisms. For some, the fact that a second-order desire is, like a first-order desire, simply a desire, cannot explain why the former, but not the latter, is able authoritatively to speak for the person. For others, a regress looms for such a hierarchical account. If we require a second-order desire in order to settle which first-order desire speaks for the agent, then the prospect of having to form a third-order desire in order to license the second-order desire seems to be in the offing. This regress, moving up the chain of desires, would eliminate any determinate fact about who we are as individuals.[45] Finally, concerns have been raised about the idea that we require an act of will in order to resolve who we are as individuals. Until the 'person' intervenes and orders and integrates her psychic life, she cannot be said to be identified with any particular element of it. But here the person is a kind of disembodied will which hovers above our mental lives, intervening from time to time, to order and integrate our psyches. Such a picture of agency is problematic because it takes as the exemplar of effective agency a situation of psychic alienation, where the person is distinct from her attitudes and views them as objects to be dealt with.[46]

---

[45] For statements of these criticisms, see Watson (1971).
[46] For this line of criticism see Smith (2000).

Frankfurt's more recent statement of his view of personhood attempts to address these criticisms. According to Frankfurt, as we've seen, being a person requires reflexive self-evaluation. This manifests itself in the formation of attitudes about our other attitudes. But on Frankfurt's considered view, a second-order desire can be identified with the agent because the agent is *satisfied* with such a desire constituting his personhood. As noted above, satisfaction consists in the *absence* of any interest in changing the attitudes in question. This absence does not require the presence of any new attitude, but merely requires that there be no movement within the agent towards a change. Satisfaction cannot be contrived or completely unreflective – acting as though we are satisfied may fool others but cannot change the fact that we are not self-satisfied. Being satisfied is similar to being relaxed – where we relax we experience the disappearance and absence of restlessness, but such a process is surely different from our contriving to appear relaxed.

This fully spelled-out account seems to address the concerns raised above. Our identification with one desire can be seen as authoritative not because it has some magical power to speak for us. Rather it reflects the fact that an agent is satisfied with her situation, invoking an absence of desire for change, providing an explanation as to why we can say that an agent should be identified with her second-order desire, an explanation that does not rely on the nature of the attitude under consideration. In addition, the threat of a regress is eliminated due to the fact that satisfaction consists in the absence of an attitude rather than the formation of a new one. Asking for a third-order desire would be redundant because there is no sense in which the agent would doubt that she is to be identified with her second-order desire.

Section 3.5 – Identification and Volitionism: A Re-Interpretation

The last objection, that identification should not consist in some act of the will on the part of the agent, cannot be easily disposed of. Recall that the concern was that requiring agency to consist in some act on the part of the agent to integrate and order her psychic life before we can identify it with her embodies an alienated picture of agency. The person, the concern goes, is not to be thought of as some distinct thing which must 'deal with' the various psychic raw material she finds within her mental life. Rather, we have a more intimate and more direct connection with our attitudes, one that is not captured by claiming that we require some distinct mental act to 'bring in' an attitude into our psychic lives and make them a part of our person. According to RR, this connection is substantiated by RA – that if an attitude can be attributed to us, insofar as it is connected to our capacities for evaluative judgement, then it can be the basis for moral appraisal. But in the first instance, all attitudes absent a special story will so qualify due to the kind of attitude that they are: judgement-sensitive. Thus there is no need to make a further theoretical move to explain how it is that we come to authoritatively possess a certain attitude. Rather, we can say in the first instance that we, *qua* agents, are "inhabiters" of our own attitudes: "They are a direct reflection of what we judge to be of value, importance, or significance" (Smith 2005: 51). Again, absent special considerations, there is the general presumption that insofar as our attitudes are connected to our capacities for evaluative judgement, they legitimately express who the agent is and open the agent up to moral assessment and appraisal. We can request reasons for these attitudes, expect an agent to defend them if challenged, and at least in the ideally rational case, expect them to be modified or extinguished if reasons are found to be lacking.

Given that these attitudes are sensitive to our rational capacities in the ways outlined in the second chapter and given that much of our actual practice of holding responsible seems to incorporate such a connection, the general presumption that is embedded in RR that we are inhabiters of our attitudes seems to be in good order.[47]

It might be thought that if an agent has a desire to go away to college as well as a desire to live in her hometown, then short of a decision on the part of the agent we cannot say where the agent stands. In one sense, if an agent has not made up her mind, then it is true that there is no fact of the matter as to where she stands. In another sense, however, there is the fact that she finds both options appealing for various reasons, and that these facts can be legitimately attributed to her person. Both elements reflect aspects of what she finds worthwhile and interesting in life, and both can be seen as expressing different aspects of her overall outlook. If after the time of decision, a person claims that the fact that she even considered staying in that small, backward town said absolutely nothing about 'who she was', I think we should not be persuaded. For while it may be part of her present *self-conception* or self-image that she is not the kind of person who would find her hometown appealing, what we are trying to determine is what can be legitimately said to be part of our *conception of the self*. And the latter notion, while importantly sensitive

[47] What does it mean to 'inhabit' one's attitudes? This is an evocative way of putting the thought that in the standard case (a case where there are no abnormalities or interventions) there is no separation between an agent and her attitudes. They are the same. Discussions of agency often emphasize that our powers of critical self-reflection create a cleavage in an agent between her brute desires and her more reflective capacities: "Reason enables a deliberating agent to step back from *anything* that might be a candidate to ground its putative requirements" (McDowell 173). But what RR holds is that this capacity does not solely manifest itself by separating a deliberating agent from her attitudes. Rather it claims that in the absence of deeply confusing or ambivalent situations, an agent's evaluative capacities have already been exercised in the very *formation* of an attitude. They provide for the very possibility of coming to acquire an attitude for which we can be considered responsible. Further, in coming to explicitly endorse an attitude (or, in noting reasons for or against this attitude), she has made it the case that her attitude reflects her various evaluative capacities and her perspective.

to a person's judgment regarding her attitudes, need not be solely determined by these mental acts.[48]

If this objection is apt, and if Frankfurt's account relies on the notion of a mental act to secure identification, then his account will not capture the relevant notion of identification that is needed to properly determine our conception of the self for the purposes of moral assessment and appraisal. So we can ask: if satisfaction depends on the fact that it 'derives' from our understanding and evaluation with how things are with us, does this derivation consist in some kind of mental act on the part of the agent? Either this reflective test for satisfaction consists in a volitional act, like a decision, or it does not.

If it does, then it seems that Frankfurt may be subject to his own criticism that our wills cannot be fashioned in the way that a sculptor fashions her product. He acknowledges that this is not a plausible picture of our relation as persons to our attitudes (Frankfurt 1999b: 100).[49] In addition, the criticism developed above that the person is

---

[48] I take the distinction between a self-conception and a conception of the self from Piper (1985). She writes: "Thus there is an important distinction to be drawn between a *self-conception* and a *conception of the self*. A self-conception picks out the basic intentional features in terms of which I actively identify myself. A conception of the self, on the other hand, provides a theoretical model that purports to explicate matters of fact regarding the nature and dynamics of the self. That I view myself as tactless is part of my self-conception; that I am in fact to be identified with my moral convictions or social relations or desires is part of a conception of the self with which I may or may not be in agreement. Thus the two are independent" (Piper 274, emphasis in original).

[49] Frankfurt claims that is incoherent to constitute one's self by an 'act of will' (a term he does not otherwise define in any extended fashion). He writes: "A person cannot make himself volitionally determinate and thereby create a truth where there was none before, merely by an 'act of will'. In other words he cannot make himself wholehearted just by a psychic movement that is fully under his immediate voluntary control" (Frankfurt 1999b: 100). While this may seem to be a repudiation of his early views of the constitution of agency via a decision, given the qualified nature of Frankfurt's claims (we cannot create volitional necessities *ex nihilo* but does this indicate anything regarding whether we can change them? Nor can we change ourselves by a movement that is 'fully' under our control – but a movement that is partially under our control could count?) it is unclear just what type of position Frankfurt means to be excluding here.

not distinct from her attitudes would also seem to apply adding a further reason to reject Frankfurt's account.

But if reflection does not consist in some volitional act, then Frankfurt has misconstrued the implications of his own argument. If satisfaction does not require a distinct mental act, but must nonetheless be 'derived' from our evaluation and understanding of how things are with us, I claim that it is not the absence of any interest in wanting to change that determines what attitudes we are to be identified with. If the idea of a *derivation* is to actually do any work in securing what it is we are to be identified with, then identification will not depend on the idea of absence that is central to the idea of self-satisfaction.[50] If anything, we are to be identified with the attitude because of the *connection* with our evaluative, cognitive and emotional capacities that go into determining how things are with us. Identification must bear this type of rational connection to our status as beings with the capacity for evaluative judgment rather than the mere fact that we do not exhibit any interest in changing how things stand with us.

Frankfurt, in attempting to avoid the idea of a 'fully voluntary act' to secure identification, has misinterpreted, I think, what is doing the real work in determining when we are to be identified with our attitudes. His claim is that we are to be identified with certain attitudes because we are satisfied with them. This satisfaction consists in an absence of any interest in changing these attitudes. But in order for this to be properly reflective of our status as beings with the capacity for reflective self-evaluation, our satisfaction must derive from this capacity. We might say that in order to be responsible for such an attitude, we must be satisfied with it in a way that reflects our self-

[50] Frankfurt elsewhere notes a connection between self-satisfaction and contentment with oneself. But this connection only reinforces the point I am trying to make. See Frankfurt (2006:17).

understanding. But locating the 'mark' of identification in an absence is problematic because it is not the idea of satisfaction that is doing the conceptual work of identification – rather it is the connection to our evaluative and cognitive capacities. Indeed, the notion of satisfaction reflects (or is consequent upon), rather than accounts for, the basis of our identification.

This opens up the possibility that our self-conception need not dictate our conception of the self. While I may have an interest in changing certain attitudes, insofar as this interest reflects certain evaluative tendencies of mine, we need not deny that we can be identified for purposes of responsibility with this attitude. From a different perspective, insofar as my evaluative and cognitive capacities imply a particular point of view, this will determine what an observer can say who a person is. It is the work of an individual's cognitive and evaluative capacities as directly reflective of the relevant parts of our agency that determines who we are, roughly our capacity to recognize and respond to reasons, rather than the individual's own judgement that is determinative of when we can be identified with our attitudes. This is not to say that the person's judgement is irrelevant – insofar as such a judgement reflects her evaluative capacities it captures an aspect of who she is. That is to say an attitude can belong to an agent with which that we are more or less satisfied.

Even if this attempt at reinterpreting Frankfurt so that his favored conception actually is a form of RR in disguise, albeit a form that runs contrary to his explicit pronouncements, is mistaken, there are still good reasons to resist Frankfurt's account. Frankfurt holds that those attitudes and actions for which we can be held responsible are one's we are satisfied with, but where this satisfaction depends upon our intellectual and

evaluational take on ourselves (or first-order desires). But while it can make sense that we require some sort of intellectual accuracy in coming to determine what attitudes we have – we may want to properly identify just what is the attitude in question – if the evaluational component is going to do any distinct work, it must go beyond the mere idea of 'getting it right'. That is, the evaluation of our attitudes involves the idea that some attitudes are right, appropriate, correct or proper to have or to reject. And as was noted in chapter two, in the ideal case, when we make such judgements regarding reasons for and against attitudes, we thereby come to form, revise or lose the attitudes that those judgements concern. Frankfurt's account does not allow, however, the evaluational capacities of an agent to do any actual work in the constitution of the attitudes for which we can be appraised and assessed, while RR gives them pride of place.[51]

It might be objected that RR is unsatisfactory here due to the simple fact that we do not often identify with our judgments about what is good or appropriate, rendering such judgements unreflective of who we are, a concern that Frankfurt's account would necessarily rule out of bounds. Gary Watson notes, for example, that "one can fail to 'identify' with one's evaluational judgements. One can in an important sense fail to value what one judges valuable" (Watson 1987a: 150). If this is an apt characterisation of how agents experience their attitudes, then it seems that RR cannot account for the fact that agents sometimes do not feel that their evaluative judgments are the best expression of who they are as persons.

---

[51] It is open to Frankfurt to simply drop the evaluational component of his view and hold that they do no work in the constitution of the self. This would, however, be contrary to his explicit commitments. It would also, I think, make his view much less plausible. It would be difficult, for example, to spell out the connection between the constitution of a person and her reasons for action, a task Frankfurt has explicitly set for himself and on which any good account of responsible agency should comment.

But this objection misconstrues the nature of the connection between our evaluative capacities and the attitudes we are to be identified with. What RR requires is that there be a connection – normative in character – between our evaluative capacities and our attitudes, in order for us to be identified with them. It does not require that the attitude somehow reflect a complete and coherent evaluative stance that the individual, on careful reflection, is prepared to wholeheartedly endorse. That is, as long as there is a normative connection between the attitude and the subject's evaluative capacities, it does not matter if she finds that she does not value what she judges to be valuable.[52] If this is taken to mean that the agent does not act upon such a judgment or finds that it leaves her cold, in some manner, insofar as both her all things considered judgment as well as her determination that 'x' leaves her cold are connected to her evaluative capacities, then they can be said to be part of her person. Thus we can accommodate the fact that we often do not identify with our value judgements by emphasizing that questions of identification are settled by the nature of the connection to our evaluative capacities, and not solely by present determinations of our self-conception. RR demands only that a rational connection to the attitude in question be present (not necessarily consciously) – and this connection can allow for the possibility that we express conflicting attitudes towards the same subject matter.

Section 3.6 – Practical Reason and the Self

I think the above shows that Frankfurt's account does not provide an adequate interpretation of RA – that is, he cannot properly account for the connection between our attitudes and the grounds for moral appraisal. Frankfurt, however, also attempts to

[52] Here again we can distinguish between our self-conception and our conception of our self – which it might be noted at this point bears some resemblance to the idea of the distinction between thick and thin versions of RR introduced in chapter two.

motivate his account of responsible agency by reflecting on the phenomenology and structure of our lives as practical reasoners. On his view, these considerations lend credence to his account of agency as satisfaction, providing evidence for his considered account of what it means to be an agent for purposes of moral appraisal. Examining such an account is worthwhile then if we are to be able to properly say whether or not RR is the best theory of responsible agency.

As we've seen individuals face a basic task as agents: they must determine what it is they care about (Frankfurt 2004: 14). The primacy of this task stems from the fact that without some determinate set of cares, an individual will lose the bearings necessary for effective choice. Without the context of choice constituted by our particular cares, an individual faces a kind of vertiginous situation – a freedom where not only can we deliberate about to do, but who to be as well. In principle, a limitless freedom provides an agent with the possibility of radically re-designing her situation as well as herself, altering her personal characteristics.

Now I think Frankfurt's claims can be interpreted in a number of ways. It bears some similarity to communitarian concerns regarding the 'emptiness' of the self in liberal political theory.[53] In that sense, just as liberal political theorists have questioned whether the goals set by a particular society ought to be authoritative for an individual, and not open to the possibility of revision, here we may question Frankfurt's claim that the particular cares that we are committed to ought to be treated as authoritative in the context of choice. While both the liberal and the objector to Frankfurt can grant that an

[53] See, for example, some remarks by Charles Taylor: "complete freedom would be a void in which nothing would be worth doing, nothing would deserve to count for anything. The self which has arrived at freedom by setting aside all external obstacles and impingements is characterless, and hence without defined purpose" (Taylor 1979: 157).

individual must care about *some* thing in order to be considered an agent and have the task of determining what to do, they claim that it does not follow that this particular care, goal or project presently guiding choice must be considered authoritative on pain of dissolution of the self. If an agent can question and revisit some of her particular commitments, then while she cannot redesign her entire self at once, as it were, she can reconsider particular commitments at different points in time.[54]

Given this line of response, I think it's best to interpret Frankfurt's claim here in light of some related remarks he makes elsewhere. He writes:

> In order for a person to be able to conceive and to initiate an inquiry in to how to live, he must have already settled upon the judgments at which the inquiry aims. Identifying the question how one should live…requires that one specify the criteria that are to be employed in evaluating various ways of living…But identifying the criteria to be employed in evaluating various ways of living is tantamount to providing an answer…for the answer to this question is that one should live in the way that best satisfies whatever criteria are to be employed for evaluating lives (Frankfurt 2004: 24-25)

There is danger of a vicious circularity here, for we can only identify how we ought to live based on evaluative criteria, but the criteria can only be meaningfully determined if we have already come to a conclusion as to what satisfies them, that is, determined what is the best way of living. Attempting to formulate and clarify the criteria that are needed to guide our inquiry can only consist in affirming the judgments at which it already aims. Determining basic criteria independently of some sense of the answer is "systematically inchoate". Call this the 'Criteria' argument.

Frankfurt's preferred solution to this circularity is to reject it as hopeless and to regard certain psychological facts of the individual as necessary starting points to our

[54] See Kymlicka (51) and following for a response to the communitarian along these lines.

practical reasoning. Without some unquestioned criteria, in the form of basic attitudes, the project of deciding what to do cannot get off the ground.

Now we should question one of the assumptions that is built into this claim, specifically, concerning the form rational justification ought to take. For Frankfurt, practical agents face a difficulty that is perhaps more familiar from debates from epistemology. In trying to determine whether our belief in 'p' is justified, we often cite other beliefs as reasons. But we can reasonably question whether our belief in these other considerations is justified in the same manner – and if an agent persists in citing other beliefs as justification, the possibility of never reaching a terminus to such an inquiry presents itself. An infinite regress is in the offing in that an agent will never be able to satisfactorily justify her belief. The only options seem to either have the chain of justification run out, the agent to circle back on one of her previous beliefs or to stop attempting to provide reasons for her belief. This difficulty has motivated various responses – what they all share is a commitment to the idea that rational justification has a foundational structure. Without some foundation upon which to base justification, the project of justifying our beliefs is perhaps not realizable.

Now whether or not foundationalism is true in epistemology (and it is far from undisputed that it may well be), assuming this structure of rational justification in the practical sphere can seem to provide illicit support for Humeans (or people of Humean inspiration like Frankfurt). For a solution to the justificatory difficulty outlined above lends support to the thought that we require some type of basic or unquestioned attitude that can ground our justificatory endeavors. In the practical realm, this amounts to positing some part of the individual as a basic given that can avoid the circularity that is

threatened when we seek to justify a particular attitude or belief. Some dialectical advantage may be gained then by having this foundationalist form of rational justification in the background. Indeed, if we assume that rational justification can only take this form, then other potential competitors are obscured from the picture, viewed as hopeless against this backdrop.[55]

Due to the background, Frankfurt's argument gains a certain plausibility that it might not otherwise have given a more open consideration of other possible forms of rational justification. Even granting this advantage, however, I want to argue that we do not require some prior set of criteria in order to be able to answer the question as to what we ought to do. We can address the 'Criteria' argument by questioning whether or not it is a necessary feature of answering the deliberative question posed above that we possess determinate criteria in the form of some basic, unquestioned attitude.[56]

A concrete example can clarify matters. Take an agent, Anne, who is uncertain as to whether or not her feelings for her partner, Bert, make it the case that she is in love with him. Part of her uncertainty stems from her sense that feelings of love are incompatible with feelings of resentment, and due to the fact that her feelings for Bert are tied up with various feelings of resentment for him, she is does not think that she can properly be said to love him. Now, suppose, after a period of reflection, Anne comes to the conclusion that feelings of love need not exclude feelings of resentment. Observing other lovers, discussing her situation with friends, taking stock of her own feelings and

---

[55] Other philosophers have made similar arguments. See, for example, Heath (1997) for the criticism that an unnoticed assumption in debates concerning practical justification is that the form that this project must take is foundational in character.

[56] Another possible interpretation of Frankfurt's argument is that a basic 'set of cares' is always sufficient for the possibility of reasoning practically. This weaker interpretation, while not obviously faithful to Frankfurt's intentions, is also undermined, I think, by the following response to the 'Criteria' argument.

actions toward Bert and reflecting on the reasons that led her to her original conclusion all help Anne see that there was a confusion in her prior thinking. The confusion lay in the thought that all genuine feelings of love necessarily preclude any feelings of resentment. Indeed, given her new vantage point, Anne can arguably make better sense of her feelings of resentment for Bert: why they are particularly hurtful for her and why she reacts so strongly towards Bert, say, given the fact that she is in love with him. Eliminating the confusion in her thought can be seen as providing a more adequate interpretation of her feelings and her interactions with her partner, a move in practical thought that consisted in part in the clarification of what her actual feelings properly entail and how they are related.

Putting the deliberative question in this way: 'do I love Bert'? forces an agent, on occasion, to take stock of her judgments and to see whether or not there is a confusion or error that lies within her thinking. Identifying such an error, and eliminating it, can be a valid step in practical reasoning, allowing for a more perspicuous understanding of a person's feelings and her reasons for action, even if Anne is mistaken at this point in time. And such gains in practical knowledge are due in part to the sort of rational reflection and examination that Anne engaged in. There is a non-accidental connection between Anne's gain in self-knowledge and the elimination of the confusion that was latent in her thought, which had caused her to rule out the possibility of loving someone she also resents.

However, if the foregoing is plausible and a legitimate move in practical thought, we should note that we needn't have made reference to any *basic* criteria in effecting this move. Anne did not have to reference any pre-determined set of criteria when she

reflected on her inchoate feelings for Bert – there was no determinate measure of the meaning of love that indicated to her whether or not it is compatible with resentment. And yet she can still note that the comparison between her prior feelings and her present understanding indicates that she has made an improvement in her self-understanding – she can make better sense of herself and her actions now that she sees the confusion that lay within her and can explicitly examine the relations of mutual support between her various attitudes.[57]

If deliberating about what to do and feel need not invoke pre-determined criteria then much of the argumentative thrust of 'Criteria' is lost. For we can grant to Frankfurt that while no agent can decide, *ex nihilo*, what is important (such an enterprise may well be systematically inchoate) this does not reduce the question 'what should I do' to the search for some pre-normative criteria that can determine what will count as important. Rather, criteria that are taken for granted can come to be seen as provisional and partial, in need of revision, and such a process of revision can take place in the absence of reference to foundational criteria. This is just as RR would have it – if an attitude is taken to be open to moral appraisal or assessment, it must be the kind of thing for which reasons can, in principle, be requested. Pre-normative criteria embedded in an agent's psychology with which they were satisfied would not be necessary to meet RR's demands.

This last point can be underlined by noting that the sharp division Frankfurt posits between the various tasks noted above (what we want and value and what we care about) is exaggerated. While Frankfurt claims that the "most basic and essential question for a

[57] The claim that comparative reasons need not rely on foundational yardsticks, as well as the slightly adapted example, are taken from (Taylor 1993).

person to raise concerning the conduct of his life cannot be the *normative* question of how he *should* live. That question can sensibly be asked only on the basis of a prior answer to the *factual* question of what he actually *does* care about" (Frankfurt 2004:26), if the above is correct, this does not follow. That is, it does not follow from the fact that we cannot formulate evaluative criteria outside or prior to engaging in the deliberation that the factual question has priority over the evaluative in grounding authoritative reasons for action.

One objection to my response to the 'Criteria' argument can be that Frankfurt's main point nonetheless stands – some kind of reference to an attitude that an agent explicitly holds will have to be made during the course of answering the deliberative question, and as a result, we will be making use of criteria in answering our self-imposed question. The use of provisional criteria, it might be objected, nonetheless remains a use of criteria, and while Frankfurt's detailed claim may require amendment, it can still stand.

This response fails however for it misconstrues the role that our attitudes and our reflection play in coming to form or adopt a new attitude. They do not function as some sort of meter stick against which we determine whether or not the options measure up – rather, they are dynamic reflections of what we take to be the reasons for and/or against holding certain doxastic and emotional responses to the situation at hand. Once they are stable, inasmuch as we consistently endorse and uphold these judgments upon reflection, they do not function as 'basic' unshakeable cares that cannot be directly questioned. Rather, they reflect, not necessarily consciously, our judgment that the considerations that went into the formation of the judgment remain good reasons. As a result, these judgments are open to clarification and elaboration, which, when pressed against the

force of specific situations and specific objections, can be directly revised and rejected if an agent determines that the balance of reasons has changed. While this process may fail from time to time, as noted above, this process is best understood as an ideal against which we measure our actual lives as practical agents.

It might also be objected that the picture that I have attempted to replace Frankfurt's with only captures a small part of our practical lives and that the kind of example outlined with Anne is an extraordinary situation out of step with how we normally and standardly settle deliberative questions. Frankfurt's account then, while not exhaustive with respect to our practical lives, captures the most important and pervasive features of those lives.

In response it must be said that the example offered seems to reflect the everyday phenomenology of practical reasoning much more faithfully than Frankfurt's. On Frankfurt's view, all reasoning bottoms out when faced with that which we cannot help but care about. And while these volitional necessities can support intermediary cares which we can make use of in practical deliberation, these intermediaries are only motivationally effective insofar as they are directly related to the foundational cares.

But the kind of practical reasoning exhibited in the example and which, I contend, is pervasive in ordinary life, need not make use of such intermediaries or the more foundational cares upon which they depend. Such reasoning points towards the considerations that bear on our choice, rather than the attitudes themselves as constituents of our reasons for action. And when the attitudes themselves come into view, they are 'transparent', as it were, to the considerations that went into their formation. That is, the considerations that counted for the formation of the attitude in the first place are the kind

of things which are brought to reflection rather than the mere fact that there is an attitude present within an agent's subjective motivational set. Everyday situations of deliberation and choice point towards whether something is a good reason or not, whether such an emotion is justified or appropriate, or whether an action is the right one considering the circumstances. These are the standard kinds of situation that call for deliberation and practical reason and they seem to be better captured by the account offered above (implicit in the 'Anne' example) rather than Frankfurt's reliance on volitional necessities. So even if criteria do sometimes play a role in our practical life, they do not play the foundational role that Frankfurt claims for them, and they are far less pervasive and important than they might seem at first blush. And it should be noted, the alternative conception of practical reason is consistent with the account of agency that is embedded within RR.

Section 3.7 – Frankfurt and the First-Person

Much of the difficulty with Frankfurt's position stems from the particular way he has framed the deliberative question. For in many of his formulations (e.g. the one just quoted which comes from the section "The Question: 'How should we live'" in Frankfurt 2004) it is posed in a third-personal vein – from an observer's point of view. Framed in such terms, the task seems in the first instance descriptive – what do we care about? – and as a result can appear insufficiently critical in spirit. Indeed Frankfurt would welcome this observation as an acceptance of the fact that we cannot reason our way to firm normative standards from the 'outside' or from the ground up.[58]

[58] This section echoes the criticisms I made above regarding Frankfurt's use of the idea of 'evaluational'.

But when the deliberative question is posed in its first-personal form, 'what should I do?' the tasks of description and critical reflection do not seem so distinct. In the example noted in the response to the criteria argument, Anne had an inchoate perception of what her feelings involved – becoming clear on just what those feelings were caused her to reflect not only on the feelings themselves, but their connections to other attitudes. In addition, we can imagine, her reflection caused her to not only re-examine the attitude itself, but it opened up the possibility of reconsidering the reasons that led her to adopt the particular interpretation of her emotions that led her to such consternation. In this manner she could re-examine whether or not the reasons which led her to believe that feelings of love necessarily exclude all feelings of resentment were in fact good ones. And such a process of reflection can inform the process of determining what it is that we are in fact feeling. Bringing these inchoate feelings into more precise view from the first-personal perspective forces the question as to whether or not there is reason to go on feeling in such a fashion – if this is in fact what I am feeling. The critical component presses itself when seen from the first person point of view, and as such, is not fully separable from the descriptive question. For an agent in the process of deliberation, the factual and normative questions are *essentially* linked. Determining *what* it is we are feeling (or believing or intending) is, in many cases, a matter of determining what *to* feel (or to believe or to intend).[59] As noted above, such a question is *transparent* with respect to the considerations that can go into determining its answer – for in answering it, we look 'through' our attitude towards the reasons which count in favour of

[59] This is not to deny that we cannot separate these questions in other contexts. For the importance of the first-personal perspective in self-knowledge and moral psychology, the 'transparency' of this perspective, as well as emphasis on the differences between the first-personal and third-personal perspectives, see Moran (2001).

the attitude itself. Such a process is characteristic of the deliberative context and is required if an agent is to answer, *for herself*, the deliberative question.

Thus the way in which the deliberative question is posed is not an option without theoretical consequence – rather, the way in which we understand the nature of that deliberative question can determine what might count as a legitimate response. Frankfurt's preferred manner of presentation obscures the essential connection between the descriptive and critical components as I have worded them by assuming that there can be a purely third-personal answer to the question that can meaningfully enter into our deliberations. But if there is in fact an essential connection between the first-personal deliberative perspective and the critical and descriptive dimensions of practical thought, then we cannot claim that a third-personal account will be satisfactory. For such an account will not capture the fact that individuals can question their commitments, revise them in light of new evidence, experiences and reflection, and at times, come to reject these commitments in lieu of new projects and concerns. Only when seen from the first-person perspective do such aspects of our practical and moral lives come into proper view.

This discussion has perhaps taken us far from where we started, given the focus on moral agency and responsibility, but it is nonetheless related. As Angela Smith has noted, the person, on Frankfurt's view, seems akin to a kind of "switch operator" (Smith 2000: 240), one whose whole personality consists in either recognizing pre-existing criteria already present within a psychology, or who is satisfied with that which they have come to find within themselves. Arguably, if my preceding arguments against Frankfurt's view of practical reasoning are correct, then this embodies an already alienated picture of

our agency. It forces us to resort to theoretical devices – such as the notion of satisfaction – that are unnecessary if we recognize that agents, insofar as they possess the evaluative capacities I outlined in chapter two, have a more direct and immediate relation to the contents of their attitudes (as I put it above, we are transparent with respect to them). Abjuring from this basic theoretical picture of an alienated view of agency is proper, I claim, because of the difficulties associated both with the idea of satisfaction and volitional necessities. RR can provide an adequate framework for responsible agency without falling prey to volitionism. Moreover, it possesses theoretical advantages over Frankfurt's account, specifically with respect to its ability to account for practices of responsibility attributions. For these reasons, while the rational relations view and much work in the area of moral agency owe much to Frankfurt's pioneering work, his considered position ought to be rejected. Whether or not the rational relations view can survive other challenges is the question I turn to in the following chapters.

Chapter Four: Responsibility – Superficial or Unfair?

Section 4.1 – Worries about RR's Conception of Responsibility

In this chapter I focus on a set of challenges to the conception of responsibility that RR specifies, the conception of responsibility outlined in RA. To recall, Responsibility as Attributability holds that:

> '(RA) An attitude or action is attributable to a person for purposes of moral responsibility if and only if it is connected to the agent in such a way that it can serve as (part of) a basis for moral appraisal'.

The challenges I will consider hold that the view of responsibility embedded in RA is either too superficial to account for the importance, seriousness and depth standardly associated with responsible agency, or that the proper way to understand the nature of responsibility is not to look to the conditions of attributability but to rather look to when it is appropriate or fair to blame a person for an attitude or action they hol. The first challenge I will call the *superficiality* problem. The second I'll term the *fairness* objection.

Before describing and responding to these objections in more detail, it would be useful to recall the discussion of Strawson's account of the nature of moral responsibility from chapter two.

Strawson's account emphasized, we saw, certain 'commonplaces'. He writes:

> The central commonplace that I want to insist on is the very great importance that we attach to the attitudes and intentions towards us of other human beings, and the great extent to which our personal feelings and reactions depend upon, or involve, our beliefs about these attitudes and intentions" (Strawson 1962: 75).

The qualities of other agent's wills, or what their attitudes or intentions are, are of fundamental importance to other agents. For the quality of will an agent displays determines whether an agent is showing goodwill or affection, and so may be liable to a

certain response from the object of those attitudes (such as gratitude), or whether an agent is expressing some form of contempt or malevolence, and so may be liable to another form of response from the object of those attitudes (such as resentment).

For my purposes, there are thus two central elements to Strawson's account of moral responsibility. First is the idea that we are part of, or involved with, or a participant in, a variety of different kinds of relationships with a variety of different individuals that help determine the conditions of moral responsibility. These ubiquitous examples of inter-personal relationships are central to his conception of moral responsibility and are, arguably, an indispensable "part of the general framework of human life" (Strawson 1962: 83).[60] They are indispensable because they embody what it is we actually care about in human life (e.g. our relationships with other people), and they allow for the expression of the reactive attitudes. It is through attention to the reactive attitudes that we determine who can be responsible in cases of interpersonal relationship. If an agent is open, in principle, to the reactive attitudes then we have a grasp on who is a legitimate candidate for a responsibility attribution.

We can discern a separate and further idea or methodological precept that is often associated with Strawson's account of the nature of moral responsibility. It is that we gain a grasp on who is an apt target of these attitudes by paying attention to actual interpersonal relationships, or more broadly, the various practices we have of holding different agents responsible. Examining just who we hold responsible and when we do so is a thus common strategy for outlining the limits and grounds of who can be responsible

[60] While I think the importance of interpersonal relationships to moral responsibility is widely held, I do not claim that Strawson's assertion that we are naturally committed to such relationships and, as a result, we cannot put these general web of relationships 'up for review' is widely held. For doubts on this score, see Wiggins (1973) and Nagel (1986), chapter 8.

(and when an agent is not to be held responsible for a certain action). As Michael McKenna presents the point, "the nature of morally responsible agency, of its conditions, is to be unpacked exclusively from the community's standards for *holding* morally responsible" (McKenna 2005: 171, emphasis in original).[61]

How does RA, and hence RR, connect with the standard understanding of responsibility that Strawson offered? In a natural and straightforward sense, RA can be seen as offering a kind of account of the nature of morally responsible agency that focuses on an agent's quality of will.[62] That is, RA seeks to understand what is it about an agent's attitudes or actions that ought to figure in a moral appraisal of those attitudes or actions. RA captures just why it is important or why we care about the quality of will of various agents – because they form the basis upon which we interact and form relationships with these individuals. Due to the basic importance of this way of relating with other individuals, those attitudes and actions which can have interpersonal significance (or alternatively, form part of a moral appraisal) should be seen to delimit the boundaries of moral responsibility. So I do think there is a natural sense in which RA is already a specification of one sense of moral responsibility that is quite common in the field.

However RA does not make the same explanatory use of the role of holding agents responsible. If anything, it seems to be silent on the role that community norms or general practices regarding holding agents responsible play in determining the conditions of moral responsibility. Rather, it holds that there is a basic connection or relationship

---

[61] For discussion as to whether this actually is an accurate interpretation of the strategy Strawson enacted and intended, see (McKenna 2005).

[62] This was a strategy that Scanlon himself, who I take is sympathetic to RR, outlined in his discussion of Strawson in his Tanner Lectures. See (Scanlon 1986).

that is required between an agent and her attitudes or actions in order to establish whether or not an agent can be legitimately held responsible. So RA seems to reverse the explanatory order that Strawson's work emphasized. RA can be seen to be more individual in character while Strawson's account more social – a contrast I will comment on below.

The two challenges that I will address in this chapter should thus be seen as two complementary attacks on RA. In a way, these two distinct challenges form a dilemma for a defender of RA (and hence for a defender of RR). For either RA is too superficial to engage in the work that it purports to do (and thus the quality of will account outlined above is not a legitimate account of moral responsibility and must be supplemented or rejected), or it misconstrues the fundamental way in which we ought to understand responsibility, which is via the notion of holding responsible (and more particularly, when it is fair to hold someone responsible) rather than attributability.

Another way to see the force of this dilemma is to examine the two kinds of responsibility that Gary Watson famously distinguished, responsibility as accountability and responsibility as self-disclosure (Watson 1996: 229). The former Watson presented as the more standard understanding of responsibility. Accountability here involves the idea that there is a certain social setting in which responsibility attributions are made, governed by various standards, one in which we place demands on others and respond (sometimes negatively) if they do not live up to these demands and standards. It is the idea that if we flout some moral obligation, an individual can hold us responsible (or accountable) for having violated the norm.[63] Thus, when an agent violates a norm of

[63] Interestingly, Watson contrasts this interpersonal element of responsibility as accountability with what he saw as the more individualistic focus of responsibility as attributability. The latter,

accountability, she deserves "adverse treatment or 'negative attitudes' in response to [her] faulty conduct" (Watson 1996: 230). However, according to Watson, we can only hold an agent accountable if an agent had a fair or reasonable opportunity to avoid committing the fault in question.

The second conception of responsibility that Watson identified, responsibility as self-disclosure, most closely parallels the kind of responsibility embedded in RA. This kind of responsibility involves an agent expressing some fundamental features of their ethical life via their words and deeds. An agent engages in an act of self-expression by adopting a way of life or course of action that communicates what it is that she stands for, her 'practical identity' or her judgements of value (similar to the judgements concerning the adequacy of reasons that were central to judgement-sensitive attitudes described in chapter one). Watson describes this conception of responsibility as consisting in adopting an end; he continues: "to commit oneself to a conception of value in this way is a way of taking responsibility. To stand for something is to take a stand, to be ready to stand up for, to defend, to affirm, to answer for" (Watson 1996: 234). Responsibility as self-disclosure is an active process whereby an agent constitutes who they are and expresses their sense of self, or "practical identity" (Watson 1996: 234) in their attitudes and actions.

However, if the charge of superficiality holds then RA will no longer be a legitimate candidate for responsibility attribution. That is, RA will be mistaken in having elevated an unimportant or thin sense of responsibility in place of the proper conception of responsibility as accountability. Presumably, the thin conception is not as important as

---

he claimed, was more interested in the relation of an individual to her behaviour rather than in the actual practice and social context of responsibility attributions (Watson 1996: 229). Below I question whether or not this contrast is apt.

responsibility as accountability because what we are most interested in with respect to moral matters is whether or not conduct like 'adverse treatment' or, at the limit, punishment is deserved. Given the complex nature that issues of self-defense, blame, and punishment by the state (to take examples that may be seen to involve adverse treatment) involve, the thought might be that what is of most moral interest is the conception of responsibility that underlies these questions, rather than questions regarding one's practical identity.

Even if the charge of superficiality fails however, if the conditions of responsibility ultimately turn on considerations of fairness (in particular, when it is fair to blame or hold an agent responsible) then responsibility as attributability will have won a pyrrhic victory, for it will have vindicated a sense of responsibility that is orthogonal to our interest in moral responsibility. Fairness considerations, it seems, are not obviously co-extensive with attributions of judgement-sensitive attitudes. It seems possible that an agent can be declared to be responsible for an action in a situation in which it is manifestly unfair to hold that agent responsible.[64] Nor, depending on our intuitions, will fairness considerations account for and vindicate the same kinds of practices of holding agents responsible. Alternatively, if responsibility is to be understood in terms of fairness, this may simply reassert the charge of superficiality in a new light, by showing how far RA falls from its intended mark.

This extended introduction to the nature of moral responsibility serves two purposes. First, it argues that the sense of responsibility that is present in RA has a prima

[64] Many preliminary reactions to the case of Robert Harris, as eloquently and sensitively described by (Watson 1987b) seem to embody the thought that while Harris is manifestly responsible (in the sense of the various attitudes being attributable to him for the purposes of moral appraisal) it is nonetheless unfair to blame him (or 'fully' blame him) for his heinous actions given the horrendous upbringing he endured as a child.

facie case for being a plausible understanding of the nature of moral responsibility (of the kind that interested Strawson) insofar as it is already present in one of the most influential accounts of compatibilist versions of moral responsibility available. Second, that if RA can respond to these two objections, then it will be the best account available of the nature of responsible agency. RA and RR will have shown to possess the resources as a theory of agency to make sense of the nature of moral responsibility and will have shown itself superior to its critics and rivals, or so I hope to show.

Section 4.2 – Is RR Superficial?The Charge Clarified

The first charge we must address then is the superficiality problem. This objection says that RA is too superficial in its attributions of responsibility and lacks the requisite depth or importance that we seek in attributions of responsibility, indeed, the kind of importance that Strawson was attempting to elucidate in noting how much interest we take in the attitudes of others. The charge that RA is too superficial has been originally and most fully developed in Susan Wolf's work. According to Wolf, theories such as RR are part of a broader set of theories which she terms 'Real Self Views' (RSV). These views hold that there is a 'real self' that is to be differentiated from other, alienated (or irrelevant) aspects of an individual's psychology and behaviour for the purposes of determining the nature of moral responsibility. There is a 'cut' to be made within our psychologies, a cut that provides the boundaries and contours of morally responsible agency.[65]

[65] Strictly speaking, it is not the fact that a 'cut' is made in an individual's psychology that renders something a Real Self View. Even the most capacious account of agency, I think, will exclude certain features of an individual's psychology from consideration. What makes the Real Self View distinctive, it seems, is the fact that the particular aspects of an individual's psychology that are excluded are viewed in some sense as alienated features of someone's psychological make-up. If this is correct, however, then many views, including RA, which I treat as a Real Self

If 'x' can be attributed to a person's real self, then 'x' is a candidate for a responsibility attribution ('x' can be held responsible for that attitude or action). If 'x' cannot be located in an agent's real self or not appropriately related to it, then an agent is able to disclaim responsibility for the attitude or action, either by being excused or exempted from responsibility as the case may be.[66] As Wolf first puts the view: the "crucial feature distinguishing unalienated from alienated action is that the will (or the choice, or the multitude of choices) of the unalienated agent arises from the agent's unalienated self – from her real self…with which the agent is to be properly identified" (Wolf 1990: 30). RSV provides the conditions of responsible agency in a way that reflects the notion of responsibility in RA.

More specifically, Wolf holds that the proper subset of the self that forms the real self is best understood as being constituted, and being governable, by an agent's whole evaluational system. What, according to Wolf, is an evaluational system? It is the collection of 'all-things-considered' judgements regarding what an agent ought to do, or what she holds to be best.[67] Thus part of John's evaluational system may be that it is best to only have one cup of coffee in the morning. It may also be part of his system that it is right and proper to attempt, as best we can, to live up to God's commandments in our

---

View in the text, would not qualify as such, for what is excluded from its purview is not necessarily an alienated part of a person's psychology. I think this instability should throw some doubt on whether or not Wolf's characterization of the Real Self View is ultimately a useful way to carve up the different positions regarding moral responsibility.

[66] I take this terminology from Wallace (1996: 118). He usefully contrasts excuses as more local 'blocks' against responsibility attributions (where the attitude displayed does not actually communicate ill will, such as ignorance) with exemptions as 'blocks' that are more global in character (such as the lack of the power of reasoning or the general immaturity of children).

[67] It is reasonable at this point to question whether this is an adequate account of what constitutes an 'evaluational system' (although Watson (1971) seems to endorse a similar view). Apart from the question of the codifiability of this system, I think it is reasonable to believe that an evaluational system may be formed by more tentative and prima facie judgements rather than well-articulated principles regarding what to do in a variety of different situations.

daily lives. Summing up this different collection constitutes an agent's evaluational system.

When an agent sees this evaluation system realized in her behaviour – and so is not subject to weakness of will or other more serious impairments of her rational capacities – then the lack of any internal or external inhibition in her action will allow her to be a rationally self-governed agent. While her evaluational system need not be actively governing the attitude or act in question – we normally do not formulate rules such as 'it is permitted to have two cups of tea in the afternoon if one feels like it' – insofar as such a judgement falls under its jurisdiction, then the core of the view is preserved. That is, the attitude would be consistent with an agent's whole evaluational system, and should the need arise, that evaluational system could trump these more local judgements (if, say, John's second cup of tea conflicted with some other feature of his psychology). An unimpeded ability to disclose one's true self constitutes the conditions of attributability for the purposes of determining moral responsibility for a RSV.

Consequently, Wolf accepts that RSVs embody the intuition present in RA and notes that "the fact that their actions [agents who are unduly constrained] are not attributable to their selves seems to justify our intuition that they are not responsible for their actions" (Wolf 1990: 34). Her attack on RSVs is thus a direct attack on RA.

Why do RSVs lack the requisite depth we require to account for attributions of moral responsibility? Wolf claims that there a variety of difficulties associated with RSVs. First, she seems to hold that the view is question-begging. She writes:

> The character of the controversy [regarding the supposed non-responsibility of agents such as severely abused children or those lacking an adequate upbringing]…may help us express more general reasons for being dissatisfied with the Real Self View. For the fact that some people are reluctant to regard the

> agents in these cases as responsible beings is enough to motivate the question of why these agents are responsible beings, if in fact they are. The reply that these agents are acting in accordance with their real selves only begs the question at this point, restating the condition that, if offered as a sufficient condition of responsibility, is itself in need of support (Wolf 1990: 39).

However, if the project begun in chapter two and continued here is at least reasonably successful, then the RA and RR will not fall prey to such a concern – for they will have independent theoretical resources to marshal in order to explain our interest in a RSV.

Wolf also offers the criticism that we often do not choose our own real-self.[68] That is, while a RSV may make us responsible for whatever acts flow from it, we may, she claims, reasonably demand that this attribution of responsibility depend, at least to some degree, on whether or not we ourselves have chosen our real self. RSVs leave out this demand and requirement on responsibility attributions, and so are incomplete as an account of the nature of responsibility, partly, it seems, due to the fact that they are primarily made from the present perspective of an agent.[69]

More specifically, RSVs are superficial because they collapse an important distinction between superficial and deep senses of responsibility. A superficial sense of responsibility is evidenced by the kinds of statements we make such as 'The wind is responsible for the tree falling'. In these statements, 'responsibility' functions to identify that something, be it an agent, or an inanimate object, is the primary cause of some further state of affairs. In this context or situation, we can identify some thing or factor as being especially noteworthy in order to signal that this thing played an important

---

[68] This concern, that we do not voluntarily choose our own characters (or upbringing or culture, etc.) dates at least to Aristotle's discussion of the voluntary in chapter three of the *Nicomachean Ethics*.

[69] Cf. Watson's remarks regarding the differences between the two kinds of responsibility above and note 63.

causal role in bringing about ‘x’, without committing ourselves to any particular account of causation. So a superficial sense of responsibility is captured by noting that ‘x’ is of note in determining the causal sequence that led to the event in question.

Now this superficial sense is obviously not what is of interest when we are elucidating the nature of moral responsibility or making responsibility attributions for morally responsible agents. I think it is reasonable for Wolf to demand that we have a conception of responsibility which does not merely note that ‘x’ played an important causal role in bringing about ‘y’. First, this would render responsibility attributions dependent upon considerations which are irrelevant from our interest in moral responsibility. If we have certain interests in determining, say, what was the second factor in a particular causal series, this may lead us to isolate and elevate one particular aspect of a series. But, I think, such a consideration (absent a very special moral context) is not something which seems to matter for questions of *moral* responsibility. Second, such attributions would lack the requisite ‘depth’ that Wolf is after. That is, given that it would not distinguish between various elements in determinations of moral responsibility (say agential and non-agential), it can be reasonably seen as superficial and eliding morally relevant differences. While Wolf holds that what provides the requisite depth is the presence of some type of choice on the part of the agent, I think her general critique is useful in indicating that there are different senses of moral responsibility at work and that an adequate account must properly characterize the moral ‘weight’ that we attach to determinations of moral responsibility.

Watson also claims that there is another sense of superficial at work in Wolf’s argument – the idea that superficial judgements of responsibility are ‘mere’ or are

"simply descriptions of a things qualities and differ in kind from moral blame in the (strict?) deep sense" (Watson 1996: 232). This aspect of superficiality has it that responsibility as self-disclosure in effect describes agents against some broader evaluative standard, in effect, grading them on some scale. So judgements of blame on this view would be akin to criticizing an individual for being bad at chess or an ungainly athlete.[70] These claims would describe an individual but would seem to lack the moral importance we seek in judgements of responsibility.

Wolf, consequently, contrasts this superficial sense (or senses) of responsibility with a deeper sense of responsibility. A deep sense of responsibility, one that is of proper concern in debates regarding the nature of morally responsible agency, holds that:

> we are regarding her [a responsible agent] as a fit subject for credit or discredit on the basis of the role she plays. When, in this context, we consider an individual worthy of blame or of praise, we are not merely judging the moral quality of the event with which the individual is so intimately associated; we are judging the moral quality of the individual herself in some more focused, noninstrumental, and seemingly more serious way (Wolf 1990: 41).

Now I should say at the outset that while there is certainly something here that is of note for debates regarding moral responsibility, it is not entirely clear just what exactly is of importance with regard to deep responsibility. Wolf claims that the mere fact that someone performed an action does not thereby settle the question as to whether or not they are 'deeply' responsible for the action. We can grant this however without coming to any further understanding as to what is the 'depth' that is lacking in superficial responsibility. Furthermore, it seems, we can grant this point, without immediately

[70] Not all philosophers see this as a problem to be overcome. For the classic account of praise and what he calls dispraise, or a theory of responsibility which sees responsibility as purely descriptive, akin to merely grading against a scale, see Smart (1961).

(unless we are begging the question against RA) seeing that a RSV must lack such a depth.

Wolf then claims that what is lacking in RSVs is that they lack a certain 'significance' with regard to their claims of responsibility. Because they cannot account for the significance we attach to individuals who are morally responsible, RA is to be rejected. Ultimately, Wolf holds that RA views (and all RSVs) lack significance because they are too individualistic in character. A proper account of morally responsible agency "must go beyond an examination of the internal complexities of a particular class of agents and consider the relation these agents, by contrast to other agents, have to the world in which they act" (Wolf 1990: 44). And this strategy leads Wolf to formulate a version of responsibility as accountability outlined by Watson, where the ability to recognize and act upon good reasons (a kind of normative competence with respect to the true and the good) is central.

It is significant that even though we can distinguish between a superficial and a deep sense of responsibility, it is not obvious, at least to me, that the content of the latter notion of responsibility is clear even if the distinction is tenable. This suggests that the proper locus of debate concerning the nature of moral responsibility occurs here – over the exact nature of 'deep' moral appraisal necessary for responsibility – and this is why it is important for RA, which purports to be offering such an account, to reject the charge of superficiality.[71] But I take it that the difficulty lies in that RSVs and RA cannot explain why it is that all deep judgements of responsibility are also judgements of RA and that their lacking this explanation is owed to the fact that they are unduly focused on the individual and her relation to her attitudes and actions rather than the broader social

[71] Compare Smith (2008: 375).

context, and the requisite conditions that accompany such a context, in which agents act.[72]

Wolf's case of Jojo, the son of a tyrannical and vicious dictator, might help illustrate this claim. Jojo is raised in a closed and rigid environment, where he is exposed to the terrible deeds his father commits on a regular basis. Since he naturally comes to see his father as a role model, Jojo himself commits such terrible deeds once he has taken over from his father. Since Jojo wholeheartedly endorses the attitudes that lead him to commit such atrocities, and when he stands back and asks 'Do I want to send this innocent person to an early death?' he finds that the answer is in fact 'yes', he can be said to be responsible for his attitudes and actions in the sense amenable to RA. Yet Wolf holds that we ought not to find Jojo responsible, as "it is unclear whether anyone with a childhood such as his could have developed into anything but the twisted and perverse sort of person that he has become" (Wolf 2003: 380). Without the proper social setting that a loving and caring childhood can provide, as well as exposure to other forms of personal interaction that are not predicated on random and capricious cruelty (with no accompanying criticism), there is no sense in which Wolf holds that Jojo is deeply responsible, even if he is responsible in the RA sense.

Section 4.3 – Is RR Superficial? The Charge Rebutted

The challenge outlined, should we grant the claim that RA is superficial in the sense that Wolf claims? To properly assess the charge, let us consider as an example of an attribution of responsibility and see, as far as we can, whether or not it is superficial in character. So, if in response to the fact that I am consistently and unapologetically late,

[72] Wolf outlines her favored account of what those conditions are, where we have a certain kind of control over who we are, in chapter six of her book.

my friend, who I've left sitting at a café waiting for me once again, says upon seeing me causally stroll up: "What a jerk!".

RA and RR would license such a claim as legitimate moral appraisals of both my conduct as well as the attitude that this conduct is reasonably seen to express. Absent an explanation or special context which might explain my consistent lateness, it seems that I display an inconsiderate attitude by not caring whether or not I arrive for our appointments on time or waste your time by repeatedly making you wait. And it seems that these attitudes are attributable to me for the purposes of moral appraisal. Could I however, in response to my friend's criticism that I am late, reply that this is a morally unimportant form of criticism, ethically superficial – a mere description of some unchanging aspect of my personality – and so not of interest to our relationship? Could I respond that criticizing my lateness would be akin to criticizing the fact that I have trouble doing higher-order math – a failing regarding my abilities but not something that speaks in any significant way regarding me as a moral individual or about my relations with other persons?

It seems to me that on the view developed here these responses would not be appropriate, that the kind of attributions that RA and RR would license would capture the deep element of responsibility that Wolf thought necessary to an adequate account and would not be analogous to saying that someone was bad at math (or failed to meet some general evaluative standard). The reason, according to RR, why such an individual would be open, in principle, to criticism, would be that both their conduct as well as the various attitudes that they possess are objectionable for the reasons mentioned above. Being consistently and unapologetically late and thus inconveniencing my friend, all

things being equal, is a form of disrespect (thus displaying a disrespectful attitude) and shows that I both do not care about them or their priorities, and it also shows that I do not care about our relationship. While in real life all things are rarely equal, the view I am advocating forms the basis upon which our other judgements are made.

Why would this kind of criticism not simply amount to noting, in a descriptive vein, that someone is lacking in some kind of respect (say, punctuality)? Why is this criticism not superficial? Because it implicates something important and personal with respect to the individual – her capacity for evaluative judgement reflected in the activity of her rational capacities. Her conduct and attitudes express and display a variety of judgement-sensitive attitudes and when placed within the larger perspective of her individual psychology we can come to see that *she* fails in properly exercising those capacities. That is, she, in the various judgements of reasons that we can attribute to her in order to make sense of her behaviour – judgements such as 'It's ok to be late', 'I'm sure they won't mind if I take a few more minutes to enjoy my morning coffee' or 'this article in the newspaper looks interesting' – shows herself to be open to the kinds of criticism that we mentioned above. And because these judgements are reflective of her rational capacities, they can then speak in a significant and deep manner about how this agent, as an individual, conducts her life. It is no idle description of her abilities but a criticism that she, as an individual, is lacking in some fashion and can provide good evidence for the moral qualities of this person (over and above any relationship to an objective evaluative standard). And so it seems without determining just what constitutes

a 'deep' judgement of responsibility or blame, surely an area of controversy, we can nonetheless rebut the charge that RA incorporates a superficial form of responsibility.[73]

This response is I think adequate to the charge as presented and can give a convincing principled explanation as to why the judgements licensed by my favored RSV (RR) are also all judgements (differing in levels of importance) of deep responsibility. They overlap just because our ability to make and form judgement-sensitive attitudes give individuals insight into our own and other person's psychologies which constitutes the evidence that we need to form moral appraisals of an individual's personality and conduct. There is thus a principled connection between this general theory of agency (RA) and the more specific theory of responsibility (RR).[74]

However, there is a further fact that we should note that helps to rebut the charge of superficiality. I'll call this fact the *interpersonal significance* of judgements about reasons (or attitudes more generally).[75] It matters greatly to us, as Strawson was keen to emphasize, what others think and believe about other individuals, how they feel and reason about their fellows. The specific way in which it matters can depend upon the particular relationship but the general point is that our relationships with other individuals, their weal and woe, depend on the content of the attitudes that the other

---

[73] Alternatively, as Rahul Kumar put the point to me, we might say that the superficial sense of responsibility that Wolf identified is just not that superficial.

[74] I say more on the nature of deep responsibility as seen by RR below.

[75] In saying this, it might appear that I am in disagreement with Smith (2008: 385). She claims that the conditions of responsibility are ones that track depth rather than the particular normative significance of an attitude or action. That is, she holds that we can be deeply responsible for an attitude or action that is nonetheless normatively insignificant. I don't mean to deny the phenomenon that Smith calls attention to, but I do think that the point made in the text is consistent with the distinction Smith draws. What I hope to emphasize is a point she herself makes that the content of our attitude can affect and potentially be grounds for various kinds of responses – and such a process seems to me to possess what I term 'interpersonal significance'. As Scanlon notes, concerning moral criticism, "moral criticism, differs from other criticism of judgement-sensitive attitudes because of the particular significance that this form of justifiability has for an agent's relations with others" (Scanlon 1998: 287).

individual holds. Now I want to claim that this fact, regarding the interpersonal significance of judgements about reasons, holds true regardless of whether or not these attitudes are expressed in words or action.[76]

We can appreciate this point by looking at an objectionable form of attitude. There are certain attitudes and forms of attitude that in themselves are open to immediate condemnation or objection, such as stereotypes regarding oppressed or marginalized groups. These kinds of attitudes are such that they immediately alter our relation with the individual that holds them. The mere fact of having an attitude is thus interpersonally significant, implicating our relations with that individual as well as the various norms (both moral and non-moral) that govern our relationship.

Regarding the negative impact of stereotypes, Lawrence Blum writes:

> Beliefs are typically part of our forms of regard for other persons. I may disrespect or do someone an injustice by thinking ill of her – for example, by seeing him as stingy, or stupid, without adequate evidence for doing so. Respect for other persons, an appreciation of others humanity and their full individuality is inconsistent with certain kinds of beliefs about them (Blum 262 - footnote omitted).

What is noteworthy here, if Blum's description of the wrong associated with the holding of a stereotype is accepted, is that the mere holding of an attitude (a particular belief in this case) is in itself damaging to a person's relations with her fellows (and especially damaging to her relations with members of the group of people that the stereotype concerns), regardless of whether or not it manifests itself in our words or deeds (e.g. whether it is directly expressed). From the content of the attitude we can infer something of great interpersonal significance – that a person's relationship with her

[76] Scanlon, citing Smith, develops the point that there is a distinction to be had between a determination that 'x' is responsible for something and expressing that judgement (Scanlon 1998: 269).

fellows can be impaired, opening her up to forms of moral criticism and response which, while perhaps unwelcome, may be the appropriate form of moral appraisal given the attitude she expresses.

What's more, Blum argues that beliefs or attitudes generally on their own can wrong in a distinct manner from the harm they may cause by their expression. He brings this out by explaining what the wrong of a stereotype is (in this instance, the stereotype that Midwesterners are unfriendly) that cannot be accounted for by the wrong of a mistaken generalization that is held in an 'unstereotypic' manner. Why is a mistaken generalization not a stereotype? Blum writes:

> Because the generalization did not shape the way I perceived Midwesterners to anything like the extent that stereotypes do. I was not inclined to see unfriendliness where it did not exist, nor to overlook friendliness where it did. My belief was not resistant to counterevidence the way stereotypic beliefs are. I did see Midwesterners as unfriendly prior to contact with them, and this was a disservice to them. But this general view did not shape my perception of Midwesterners with whom I came in contact (Blum 263 - footnote omitted)

This explanation of the difference between a mistaken generalization and a stereotype is useful in that it both underlies the interpersonal significance of the attitude itself (over and above any purely descriptive or informational mistake that an individual may have made), thus showing it to be a worthy candidate for moral appraisal, and it also reinforces the holistic character of our judgements of attributions of responsibility, a character I emphasized in chapter two in developing RR. The shaping functions of our judgements, their connections with the other attitudes that we hold as well as with what we perceive to be of significance, or reasons for action, all provide further evidence for thinking that the attitudes in question are of interpersonal significance because they implicate a broad swath of a person's psychology and are an accurate and appropriate

reflection of her various views on a subject. The idea that this attitude is enmeshed within a person's psychological web and is tied to other aspects of her personality, most notably, her rational capacities, seems to underline that it is a true reflection of her self.

Consequently, I think that we can say that RA is not subject to the superficiality problem. Before moving on to consider whether or not the fairness objection is successful against RA, I want to comment briefly on the idea of 'deep' responsibility that is favored by RA.

Recall that deep responsibility for Wolf was a judgement that reflected something important or morally serious about an individual. According to Angela Smith, RA and RR can account for the depth requirement on judgements of responsibility by noting two facts. First that "moral criticism, unlike many other forms of negative assessment, seems to imply something about our *activity* as rational agents" (Smith 2008: 380). As noted above, our rational activity consists in the way in which we recognize and respond to various kinds of reasons and enact those considerations in our conduct. When our attitudes or our actions come up for criticism, as Smith says, we can see how this criticism implies something about our activity as rational agents.

The second fact that Smith notes is that moral criticism "by its very nature, seems to address a *demand* to its target" (Smith 2008: 381, emphasis in original). When we are the recipients of moral criticism, we can understand such criticism as demanding that we justify our conduct, disclaim that we actually committed the conduct in question or explain ourselves and acknowledge the fact that we've committed a wrong of some kind. Smith understands such a demand as making an "implicit demand for justification" (Smith 208: 381) to the agent who has held an attitude or performed a certain act. Moral

criticism thus offers a challenge to an agent, a challenge that implicates them as a moral agent, because such criticism speaks about an agent's activity as an agent. Such a challenge invites a response. While the agent who is criticized may not care or take any interest in such a challenge, the in principle demand nonetheless applies. It amounts to a kind of standing obligation moral agents have to either justify their attitudes and actions or modify them in light of the criticism that is offered.

Two points about this account of deep responsibility are of note. First, this account of deep responsibility is intended to differentiate it from a superficial account of responsibility. According to Smith, the implicit demand for justification regarding an activity of the rational agent is what differentiates criticism such as 'You're a jerk' from unwelcome description of an agent, such as being called ungainly. This is because such criticism invites the individual to either justify their conduct or to alter it in light of the criticism.[77]

Second, this account of deep responsibility provides the needed principled basis to show why it is that the type of agency that RA implicates is co-extensive with judgements of deep responsibility. As Smith writes:

> These features of moral appraisal – the fact that it implies something about a person's activity as a moral agent, and the fact that it addresses a justificatory demand to its target – together imply that it is appropriately directed only at features of a person that can be said to reflect her practical agency (Smith 2008: 381).

According to RR, those features of a person which reflect a person's practical agency are her judgements about reasons and are thus attributable to her for the purposes of moral appraisal. As previously noted, there is thus a response to Wolf's claim that it

[77] As Smith writes: "Criticism, in this case, is not mere unwelcome description, but calls upon a person to re-evaluate the grounds of her attitudes and intentions and to modify them if those grounds seem faulty or insufficient" (Smith 2008: 386).

would be an unexplained accident why a RSV and judgements of deep responsibility would overlap. They do overlap because those aspects of a person's real self are attributable to her and she is thus deeply responsible for those elements. The needed principled basis has, on this account of RA, been secured.

I think then that both RA and RR have the resources to adequately address the superficiality problem that Wolf raised. The second problem, however, remains. This objection claimed that we should look to our intuitions regarding when it is appropriate or fair to hold an individual responsible in order to determine the grounds and nature of moral responsibility in general. That is, their objection proposes an explanatory strategy that is the reverse of the one that is so far been pursued here, where we look to see whether an attitude or action is attributable to an agent for the purposes of moral appraisal, and not to whether or not it would be appropriate to hold the agent responsible for the attitude or action in question.

Section 4.4 – Is RR Unfair? The Charge Clarified

The fairness objection states that we only hold individuals responsible for objectionable attitudes and conduct when it is in fact fair to do so.[78] One way of putting the point has already been seen in Watson's discussion of responsibility as accountability. On this view, because responsibility as accountability raises questions of 'demands' on

[78] One might wonder whether or not there are good grounds to consider the fairness objection as it might seem to be orthogonal to the question of responsibility as attributability, especially given the distinction between being responsible and expressions of that determination that defenders of RR adopt. That said, I do think it is worth considering the challenge. It seems natural for many to resort to intuitions regarding the appropriateness of holding an agent responsible in setting out an account of moral responsibility. Further, it seems possible to interpret the fairness strategy as offering an account of responsibility as attributability: an attitude is only attributable to an agent for the purposes of moral responsibility if and only if it would be fair to hold that agent responsible. While such an account risks circularity, it could be seen as 'reading off' the conditions of moral responsibility from our intuitions of when it is fair in fact to hold some accountable.

individuals, the burden of these demands generate requirements of fairness in order for an agent to be held (and thus be) responsible. As Watson writes, "[i]t is these concerns about fairness that underlie the requirement of control (or avoidability) as a condition of moral accountability" (Watson 1996: 235). According to Watson, there is a sense of responsibility where if an agent acts or conducts themselves in a certain fashion then they will be open to various forms of adverse treatment. However, it would appear to be wrong, or unfair, to engage in this adverse treatment if an agent did not have adequate opportunity to avoid the action in question. When we make a demand to an agent to justify or explain their conduct, there is the possibility of being treated adversely. Because this is the form that the demand generally takes, this explains why it is that it concerns fairness. Unless agents have adequate opportunity to avoid performing certain actions, and they should not be open to what Watson terms "sanction" (Watson 1996: 237). This is because, according to Watson, sanctions consist in a form of adverse treatment, treatment that is unwelcome by the individual in question, treatment that goes beyond merely noting that an individual has an objectionable attitude to actually treating an individual in an unwelcome and unfriendly fashion.[79] The propriety of adverse treatment explains why it is that accountability attributions are governed by norm of fairness.

Similarly, Wallace writes that the "conditions of responsibility are to be construed as conditions that make it fair to adopt the stance of holding people responsible" (Wallace 1996: 15). When these conditions of fairness indicate that it would not be appropriate to hold an individual responsible, this would be equivalent to saying that the individual in

[79] Watson writes: "One's blaming attitudes are unfair if it would be unfair for whatever reason to subject others to the adverse treatment to which one's attitudes dispose one" (Watson 1996: 239).

question is not in fact responsible for the conduct in question (either on this particular occasion or on all such occasions). This is what Wallace terms a "normative interpretation" (Wallace 1996: 1) of the free will debate. Why is it, we may ask, that considerations of fairness provide insight into the conditions of moral responsibility? What is the connection between the two?

Wallace argues that in order to properly make sense of the disagreement between compatibilists and incompatibilists we must resort to the fairness interpretation of moral responsibility. According to Wallace, the specific concern raised by incompatibilists is that it would be inappropriate in some sense to hold individuals responsible, in every single instance, if determinism were in fact true (Wallace 1996: 90). (Presumably what is inappropriate here is just the fact that it is unfair). Given this is the form their concern takes, then it seems eminently sensible to interpret the disagreement between the two camps as a disagreement over the norms which govern our responsibility attributions. In other words, it is a disagreement over when it is legitimate to hold an individual responsible – and from these norms we can 'read off', as it were, the conditions of responsible agency (as well as help resolve, as much as we can, the debate between compatibilists and incompatibilists).

Another reason that Wallace adopts this strategy is that he cannot "see how to make sense of the idea of some prior and independent realm of moral responsibility facts" (Wallace 1996: 88). Wallace is sceptical that we have any grasp of how to make sense of facts regarding the nature of responsibility that is independent of "our activities and interests in holding people responsible" (Wallace 1996: 88). While he presents no argument for this view, he claims the burden of proof lies with his opponent (especially

the opponent who construes these facts as metaphysical in character). So there is evident motivation in construing the nature of responsibility in terms of the norms governing when it is appropriate to hold individuals responsible.

Section 4.5 – Is RR Unfair? The Charge Rebutted Directly and Indirectly

There are two responses (one direct and one indirect) available to RA that rebut the strategy that Watson and Wallace outline.

Before addressing these responses, we ought to consider what seems to be a quick and, if accurate, devastating response to the fairness strategy. As John Martin Fischer and Mark Ravizza note, regarding the fairness strategy that seems to be embedded in Strawson's theory,

> 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. Surely it is possible that one can be held responsible even though one in fact is not responsible, and conversely that one can be responsible even though one is actually not treated as a responsible agent. By understanding responsibility primarily in terms of our actual practices of adopting or not adopting certain attitudes towards agents Strawson's theory risks blurring the difference between these two issues (Fischer and Ravizza 18 – their emphasis)

While this point is well taken, it seems to beg the question against the Strawsonian. Both Strawson and his opponents ought to acknowledge that there can be a gap between an agent who is responsible and an agent is held responsible – so an adult who inappropriately holds a child responsible for some minor infraction ought to be viewed as making a mistake and an individual who patronizingly abjures from holding an adult responsible for a significant slight ought to also be seen as making a mistake. Given that these possibilities are acknowledged on all sides, it is hasty to convict the Strawsonian of simply ignoring this fundamental distinction.

Rather, if this objection is to be seen as plausible, it must posit a principled divergence between our intuitions regarding when an agent is responsible and when an agent is held responsible. However, the way in which such a divergence would be made evident would be to see whether or not the fairness strategy that is adopted by thinkers such as Wallace is successful in making sense of our practices of responsibility attributions. At this stage, then, Fischer and Ravizza's claim is premature. We require a more detailed account of the nature of the divergence between being and holding responsible before we can properly accuse Strawson and his followers of ignoring a fundamental distinction of moral responsibility.

The first direct response to the fairness strategy claims using the norms of holding responsible to explain being responsible conflates a number of distinct stages in our judgements of responsibility. Angela Smith helpfully offers the following typology for distinguishing the various sense of the phrase 'A holds B responsible for X': "(1) A holds B to be responsible for X (in the sense of 'open' to moral appraisal for it); (2) A holds B to be culpable for X (in the sense of 'open' to legitimate moral criticism for it); or (3) A blames B for X" (Smith 2007: 469). Each number represents a distinct moment, on this view, that demands different sets of considerations. At the first stage, Smith argues that the considerations which are appropriate to determining whether or not an agent is open to moral appraisal at all turn on whether or not an agent's attitudes and actions are properly reflective of her rational capacities. According to Smith, it is at this stage that the conditions of moral responsibility come in to proper view.

Some support for this contention comes from the fact that we can ask whether or not an agent is responsible for an act prior to making any determinations as to whether or

not it would be appropriate to respond to the act in any way. So, in coming to determine who in fact took the cookies, we may ask whether or not it was Mary or Richard who took the cookies without yet having made any determination as to whether or not any particular response is appropriate in this particular instance. But answering the question 'who took the cookies?' seems to be an answer to the question who is responsible for 'x' and thus seems to concern the nature and conditions of morally responsible action. Thus, Smith claims, this question, regarding the nature of morally responsible agency, need not be answered through reflection on when it is legitimate to hold an individual responsible.[80]

At the second stage, once we have established that an individual did in fact qualify as a responsible agent who could be morally assessed for the conduct in question, we may offer criticism of the attitude or action in question if it fails to meet some type of relevant moral standard. We can engage in moral criticism (if we have good reason to make such criticisms).

Before proceeding to the third stage, I think we should note that senses (1) and (2) of holding responsible are more closely tied than Smith seems to suggest in this particular article (Smith 2007). Indeed, given her presentation of the reason for the principled connection between the theory of agency expressed in RR and the judgements of deep responsibility outlined above, she herself, I think, ought to hold that the conditions of responsibility should not solely be limited to sense (1) but must also include the considerations relevant to sense (2). This is because in order to make sense of moral criticism (which, I take it, is shorthand for the broader discussion of moral responsibility)

[80] I don't mean to suggest by this that the question 'who took the cookies' is merely a question of determination of fact, although it is that. It is also a question as to who would be a legitimate candidate for moral appraisal for the act, in addition to having brought it about.

we must allow for both the sense in which we attribute attitudes to agents as well as the nature of the justificatory demand that criticism, at least in principle, poses to agents. So the conditions of responsibility will be sensitive to considerations from senses (1) and (2). But this is no objection to the overall strategy that Smith forwards. It is not fatal because on Smith's view, one that RR shares, there is a principled reason for connecting senses (1) and (2) which allows to explain the nature and force of judgements of moral responsibility from either direction, as they both ultimately have the same target, an individual's judgement-sensitive attitudes. While Smith may object that we can make responsibility judgements regarding attitudes and actions that are of little or no significance (e.g. 'I feel like scratching my hair'), this issue is a red-herring. Our attitudes and actions are of general interpersonal significance and as a result can qualify as a basis for moral appraisal of an agent. Scratching one's hair is not normally of moral significance – but it is not impossible that it should be so, depending on the case.

The third stage is a stage of active blame, where we have to determine whether or not, given that we have determined an agent both to be responsible for the action (in sense (1)) and culpable (that is, liable to moral criticism in sense (2)), to direct feelings of resentment and contempt towards the individual in question. The question at this stage becomes "whether it would be appropriate to take up attitudes of anger, resentment, or indignation toward her, or to actively express these attitudes in any way" (Smith 2007: 471). Smith claims that this question is importantly distinct from the first two because it calls upon particular individuals to respond to other specific persons in a certain way. So if my brother steals my freshly baked cookies, the fact that the thief in question is my brother can be seen to be a relevant consideration in determining whether or not certain

kinds of feelings would be appropriate to have or to express towards him. Moreover, Smith writes that "this question [whether sense 3 is appropriate] is about how we should respond to *the person as a whole*…we may need to take into consideration things other than her responsibility and culpability for her action or attitude in determining the appropriate response" (Smith 2007: 471), considerations such as whether or not there is a family relation or past history of interaction.[81]

For our purposes, Smith's claim that we can make good conceptual sense of (1) and (2) without having to resort to considerations appropriate to sense (3) seem to show that we are not required to 'read off' the conditions of responsible agency from whether or not it is appropriate to hold an agent responsible. Indeed, if considerations that are irrelevant to attributions of responsibility are relevant to determining whether or not blame is appropriate in a particular set of circumstances – say the fact that there is a past history of victimization with respect to an individual or group, which counts against expressing blaming attitudes – then the Wallace/Watson strategy will be directly rebutted. The direct response thus provides some principled reasons for thinking that there is in fact no constitutive relationship between being responsible and holding responsible, or that there may be a systematic gap between our judgements of the two. As a result, we have good grounds to reject the methodological precept embodied in the fairness strategy.

The indirect strategy accepts part of Wallace's proposal – that determining when someone is to be held responsible is a normative matter and requires the use of first-order moral theorizing. But rather than coming to endorse the particular form of moral theorizing that leads Wallace to hold that the conditions of responsibility are to be read

[81] Smith also surveys other kinds of consideration such as standing, the agent's response as well as the significance of the fault that determine whether or not the 'active expression' of the reactive attitudes would be appropriate. See Smith (2007: 478-483).

off the conditions of legitimately holding an agent responsible, this theorizing leads to a version of RR. Take Watson's general concern that unless an agent has an adequate opportunity to avoid certain actions, then it would be unfair to hold them accountable for such an action. Watson holds that unless it is possible to avoid incurring a certain obligation, or, once incurred, unless one has the ability to comply with the obligation, then it would be unfair to hold an agent responsible and subject them to sanction for performing (or not performing) a certain act, as the case may be.

This is certainly a first-order moral judgement – Watson provides the case of a hijacker demanding of a hostage: "'I'm holding you responsible for keeping the others in line. If they try anything, you'll answer to me'" (Watson 1996: 237). He claims that this demand is unfair both because the individual (the hijacker) does not have the authority to make it as well as the fact that the person to whom it is addressed may not possess the ability to comply with it. As a result, it places an undue burden on an individual, something for which it would be inappropriate to hold them responsible.

But the indirect response can either reject that this is an appropriate first-order moral judgement, or alternatively, accept the moral judgement, but note that this first-order moral theorizing is separate from determinations of moral responsibility. Indeed, as I note below, such a fact is presupposed by the disagreement over the first-order moral theorizing.

With respect to the first strategy, of accepting the appropriateness of a first-order moral judgement, it is sometimes said of agents who cannot comply with certain norms (e.g. psychopaths) that they are as a result not to be held responsible for violating those same norms. While much can be said as to what sense psychopaths 'cannot' comply with

moral requirements, for my purposes, discussions of the responsibility of psychopaths seem to frequently stem from first-order moral disagreements, with the concerns regarding the conditions of responsibility following in its train. If the moral judgement however leads one to endorse a form of responsibility attribution that is consistent with RR, then we have indirect evidence that our practice of holding agents responsible is already committed to a kind of responsibility that is not hostage to facts regarding the practice of holding agents responsible.[82] All this to say is that some moral judgements seem consistent with RR, and better made sense of by RR, as Adams (1985) and many others noted have noted. Even accepting the fairness strategy then is not necessarily a barrier to the thought that RR is our preferred theory of responsible agency.

With respect to the second concern, that our moral theorizing determines our account of responsibility, recall that it held that an agent did not have adequate opportunity to be able to enact a particular demand entailed that it would be inappropriate to hold the agent responsible for not living up to the demand. But the propriety of such a judgement, it seems, depends on the details of our favored moral theory. Actual ability to do 'x' sometimes seems to be morally relevant, depending on the context, but its larger moral relevance obviously depends on important and controversial philosophical issues, such as whether or not 'ought implies can' and the nature of what we can reasonably request of others. But then it is open to a moral theorist to say that "the claim that an action is wrong should not be identified with the claim that imposing the 'burden' of

[82] Psychopaths have, in recent years, become somewhat of a hot topic in philosophical discussions of responsibility (as well as motivational judgement internalism, moral rationalism and sentimentalism as well as other areas of moral theorizing). While my brief comments do not do justice to this literature here, the probative value of the particular features of psychopaths' psychological makeup (a makeup which is by no means universally agreed upon – see Borg and Sinnott-Armstrong (2013) for doubts on this score) is not immediately obvious from a theoretical perspective.

adverse moral treatment on the subject is justified" (Scanlon 1998: 286). On such an account, the supposed costs associated with being blamed would be held separate from an account of what it is that determines whether or not an act is wrong (and thus providing our first-order moral theory). If it is intelligible to separate our first-order moral theorizing from our judgements of when it is appropriate to hold an individual responsible, and thus to separate our judgements of responsibility from our expressions of those same judgements, then we have evidence again that the conditions of being responsible are distinct from the conditions of holding responsible.

More substantively, it is open to dispute just what is the nature of the 'burden' of a responsibility attribution in question that motivates Watson's concern. While many, Watson included, interpret a judgment that 'action x' falls under a certain moral demand (and is thus responsible for living up to that demand) as a kind of sanction (akin to adverse treatment), this need not be viewed as the only account available. If there is an alternative account available of the 'force' or 'pressure' that one legitimately feels when moral demands are placed on them, then such a theory of the force of moral criticism which does not interpret it as a kind of sanction could supply the needed account without resorting to the stance of the moral judge in determining how it is we hold various agents' responsible. In this case, the account of moral criticism (or deep responsibility judgements) outlined above would fill this need.

Both the direct and indirect strategies tend to converge – there are conceptual as well as first-order reasons for preferring the strategy where by examining the quality of an agent's will and the significance of the attitude or action in question we can offer an adequate and theoretically satisfying account of responsibility that does not resort to

using our practices of responsibility attribution, in particular, our intuitions regarding the propriety of fairness, to determine the grounds of responsibility.

To recapitulate briefly: in this chapter I have attempted to vindicate the conception of responsibility as attributability that is embedded in RR. In rebutting concerns regarding superficiality and fairness, RR has again displayed I think the fact that it possess resources both as a philosophical theory about responsibility and as meshing with attractive first-order moral theories. In the following chapter, I turn to questions of practical rationality, in particular, akratic action, in order to see whether or not RR can make sense of responsibility in these cases.

Chapter Five – The Value of Moral Responsibility

Section 5.1 – The Value of Moral Responsibility

What is the value of responsible agency?[83] I think this is as important a question as the question of the nature of responsibility itself. Harry Frankfurt has identified an additional condition on an adequate theory of freedom of the will: "another condition that must be met by any such theory, by making it apparent why any such freedom of the will should regarded as desirable" (Frankfurt 1988b: 22).[84] Changing the focus from a theory of freedom of the will to conceptions of responsibility – a move that Frankfurt himself might endorse – John Martin Fischer has put the question as: what is "the value of acting so as to be morally responsible" (Fischer 2006:118)?[85]

There are at several reasons why we should be concerned with elucidating the value of moral responsibility. Much of the disagreement canvassed in the previous

---

[83] Since RR is a theory of agency, it might seem appropriate to frame my question in terms of the value of responsible agency rather than the value of morally responsible agency, as the latter encompasses more than the former – whatever value I exhibit in enjoying the pleasant taste of orange juice will be distinct from the value that inheres in instances of morally responsible acts of agency. That said, the way in which I understand my project in this chapter is to outline a theory of the value of responsible agency with a view to how it would inform an account of the value of morally responsible agency. Given that an account of responsible agency is a necessary condition for a full account of the nature of moral responsibility (at least as how I outline responsibility as attributability in chapter two), the value of the former should shed light on the value of the latter. As a result, I will treat the two as interchangeable, without claiming that they should be identified, or that there are no other relevant differences between the two concepts. Thus, my question is: what is the value of instances of responsible agency (e.g. actions) that render an agent a suitable target of moral appraisal and criticism?

[84] Without such an account we would not be able to explain what is so bad or frustrating in regards to a person who is a helpless, passive bystander to forces that move her to action, forces that operate without her own involvement.

[85] I will try to remain as agnostic as is possible as to the exact nature of moral responsibility, given that is partly the topic at issue, to avoid begging any questions. Clearly, my favored account is now established. To repeat a point made above, P.F. Strawson's account of the reactive attitudes is often taken to be a paradigmatic understanding of the nature of responsibility, and can serve as a kind of neutral understanding. For the purposes of this chapter, then, we can understand morally responsible agency as being an appropriate target for the reactive attitudes, or being open, in principle, to the Strawsonian reactive attitudes (Strawson 1962).

chapter turns, I think, on different conceptions of the importance (or danger) of being held responsible for some act or attitude. Many of the intuitions that drive debates regarding the superficiality of conceptions of responsibility or debates concerning whether or not being responsible is conceptually prior to holding responsible are intuitions that stem from the role that responsibility is meant to play in our lives. In other words, these debates are driven, at least in part, by disagreements over the value of moral responsibility. Trying to set out what this value is has the virtue I think of isolating just where principled disagreements actually lie, especially for compatibilists (who will be the focus of the chapter).

More broadly, as Fischer emphasizes, one source of disagreement between compatibilists and incompatibilists concerns the value of moral responsibility (and its associated practices) rather than about more familiar issues concerning the nature of agent-causation or conditional analyses of 'can'. If realizing the value of moral responsibility requires some type of deep or ultimate control, then this would provide some preliminary evidence for participants in the free will debate.[86] Given that some of the disagreement between compatibilists and incompatibilists stems from differing conceptions of the value of responsible action, bringing this into the open should clarify and help advance the debate.[87]

---

[86] From an incompatibilist perspective, Strawson (1994) expresses scepticism regarding the possibility of moral responsibility that seems to depend on the idea that moral responsibility ought to play a certain kind of role in our lives (and so ought to realize a certain kind of value). See, in particular, premises 3 and 4 of the 'Basic Argument' for this view. From a compatibilist perspective, Scanlon (2013) has recently claimed that a proper account of desert (centrally related to the idea of the value of moral responsibility) would also shed light on the debates between compatibilists and incompatibilists.

[87] See Fischer (1994: 83-85) for the characterisation of the free will debate as beset by a kind of 'dialectical stalemate', where progress is, if anything, piecemeal and hard-won.

Finally, such a project helps rebut current critics of the role that moral responsibility plays in our lives. Some philosophers view attributions of moral responsibility as a pernicious force, a remnant of a primitive and brutal past that is better abandoned and forgotten than celebrated and extoled. Bruce Waller's *Against Moral Responsibility* is exemplary in this regard. Given a mature and complete naturalistic understanding of the world, one that takes seriously the developments of natural sciences, evolutionary biology and accepts determinism but does not hold out hope for a libertarian vision of free will, one must reject a morality that includes moral responsibility. Waller claims: "just deserts and moral responsibility are atavistic remainders from the age of mysteries and miracles, and contemporary naturalists should eliminate them altogether" (Waller 140).[88] These remainders are actively harmful and like any injurious or destructive practice, they ought to be abolished. Our lives would be measurably improved, according to Waller, if we were to rid ourselves of our propensity to judge that individuals are responsible and to structure our relations with and reactions to others around those judgements.

Now while libertarians may be thought to have a traditional insight into the value of moral responsibility[89], I want to argue here that compatibilism has ample, and largely untapped, resources to develop such an account.[90] Compatibilisists claim that while it

---

[88] Waller's insistent rejection of moral responsibility is surely linked to his thought that it is best understood as a form of 'retributivism', connecting 'being responsible' to the propriety of reward and punishment. But the sentiment of rejection is not new, nor limited to incompatibilists. See Smart (1961) for an earlier version of such scepticism.

[89] See Kane (1998), chapter 6, for a libertarian account of the significance of free will.

[90] Various compatibilists have offered differing accounts of the nature and structure of responsibility, but have not, to my mind, adequately accounted for the distinct value of moral responsibility. So, as noted above, while Susan Wolf criticizes 'real-self' views as being shallow and superficial (and thus lacking the deep and pervasive character necessary for responsible agency), the *value* of this 'depth' requirement remains under-developed (Wolf 1990: 35-45).

may well be true that all events are causally determined, this should have no negative impact on how we go about leading our lives.[91] We need not make any radical revisions to the way we live as responsible agents even though all our actions and thoughts may be determined.[92] Such a position should be able to adequately account for the distinct value of morally responsible agency.

This type of position has been developed in John Martin Fischer's work. He offers a detailed proposal for a compatibilist account of the value of moral responsibility.[93] While Fisher's proposal is attractive, it is insufficient to account for the value, a point I illustrate with a number of counter examples. Moreover, it does not adequately capture the interpersonal element of the value that I hold is central to a proper account. I close by sketching an alternative account of the value that emphasizes the ideal of justifiability as capturing what is significant about the value of moral responsibility.

Similarly, while Gary Watson (1996) notes a bifurcation in our understanding of responsibility and sketches their distinct sources, his ultimate concern is to vindicate that responsibility as self-disclosure is a legitimate form of responsibility, even if not as fundamentally important as responsibility as accountability. What is the overall importance of responsibility is not directly addressed. Moreover, the brief account of the value of responsibility as self-disclosure that he does offer mirrors to a large degree the one criticized here (Watson 1996: 231-234).

[91] Compatibilism, generally, holds that while we may be determined in various ways we can nonetheless possess a free will and be held responsible for our attitudes and actions. This is an obviously loose definition of what constitutes compatibilism, but it will serve for present purposes. For a more expansive and detailed survey of the various features of the view, see Watson (1987a).

[92] Not all philosophers accept this claim though. See Vargas (2005) for a philosopher (although not, strictly speaking, a compatibilist) who argues in favour of a 'revisionist' account of moral responsibility.

[93] I should note that what primarily motivates Fischer to introduce his proposal is his interest in vindicating his favored theory of responsibility. He wants to show that agents only require guidance control in order to be morally responsible for their actions, rather than requiring a more robust kind of control, regulative control. He associates the value of making a difference in the world with regulative control, while guidance control exhibits the value of self-expression. But, Fischer claims, the value inherent in moral responsibility is self-expression, rather than difference-making, thus showing why we ought to be think that guidance control is sufficient for moral responsibility rather than the more robust regulative control. Since these concerns are internal to Fischer's own theory, and I think the question of the value of morally responsible agency is independently interesting, I leave them aside. For Fischer's distinctions, see Fischer and Ravizza (1998), *passim*.

Section 5.2 – Fischer's Account of the Value of Moral Responsibility

So what is it on a compatibilist view that is worthwhile about being responsible?[94] Fischer claims that "the 'value' of morally responsible action is understood as analogous to the value of artistic self-expression" (Fischer 2006: 114). This is an analogy – Fischer is not comparing individuals, as some existentialists perhaps did, to a sculptor that can shape their life *ex nihilo,* or from whole cloth.

Fischer claims that the value of intentional action comes from the fact that morally responsible behaviour expresses who we are in a certain way. We open our self up to the world in a manner that discloses some part or aspect of our personality. This process of opening up, however, is not to be construed as a piece-meal operation, where we take one instance of self-expression and add it on to a previous, perhaps in anticipation of some further instance of self-expressive action. Rather, there is a connected whole into which our self-expressive behaviour fits – there is a narrative structure that extends into the past and into the future in which our self-expressive actions fall. This, in a way that is similar to artistic creativity, is what is important about morally responsible action. Citing David Velleman's work on the nature of well-being and time, Fischer holds that the meaning of a bit of expressive activity only arises from its place in a narrative structure.[95]

Fischer thus claims that "in performing an action at a given time, we can be understood as writing a sentence in the book of our lives" (Fischer 2006: 116). Our lives are narratives which extend both into the past and into the future – we are the authors of

[94] One requirement that I simply note in passing on such an account is that it should make sense of the worth of both the status associated with being a responsible agent as well as the various exercises of that agency. While these senses are obviously related, perhaps even constitutively as Raz (2001) would have it, both must be accounted for in order for the theory to be adequate.
[95] Velleman (1991).

this book in the relevant sense when we contribute sentences to it that structure and shape the whole. So an agent in acting 'writes a sentence' whose "meaning is fixed in part by relationships to other sentences in this book, that is, by the overall narrative structure of the life. In acting…[an agent] writes part of the book of his life and gets its meaning from its place in this story" (Fischer 2006: 116).

So, self-expressive activity is valuable.[96] The value of this expression does not stem from whatever it is that we may be doing at the time; nor does it stem from whatever it is that we may produce or whatever difference it is that we make to the structure or direction of the world. Indeed, the value of self-expressive activity does not lie at all with the content of that self-expressive act (even if the bits of activity are propositions, if we are to read Fischer literally). While there are important relationships between a self-expressive act and other aspects of our life, the importance of those connections does not lie in the fact that the self-expressive act discloses what an agent judges to be good or of value. The value of responsible action comes from the place in which this particular piece of action fits within the larger narrative structure of our lives.[97]

Compare this account to the situation of an artist making a sculpture. The sculpture may be good or bad, interesting or boring, the sculpture may even be

---

[96] I don't think that Fischer is claiming that self-expressive activity is 'intrinsically valuable', in the sense that the value can be specified by referencing only non-relational properties of the thing in question, but rather that it possesses what Korsgaard calls 'final value' or the value something has for its own sake. For the distinction between intrinsic and extrinsic value as contrasted with the distinction between final value and instrumental value, see Korsgaard (1996a).

[97] A further concern for Fischer's account of *morally* responsible agency is that it seems the above formulation would permit an amoralist to recognize the value of responsible action. But it should not be possible for an amoralist to affirm the value of morally responsible agency. Thus perhaps Fischer's account is best viewed as a theory of the value of responsible agency *simpliciter*.

incomplete or defective in some other way, and yet we can still ask, Fischer claims, what is the value of the creative activity of the sculptor herself throughout this process? The value of the creative activity is similar to the value that morally responsible action contributes to our lives, in that it allows us to contribute to an evolving and extended dramatic narrative. It does not entail the kind of ultimate control some incompatibilists and some compatibilists seek, similar to the control an omniscient narrator might possess, but it does suggest that our expression matters, in some sense, to what the meaning of the various parts of the story will be. Our self-expression is necessary (and even sufficient if we note that even bad sculptures are expressions of valuable artistic activity) to constitute a valuable instance of morally responsible agency.

I think this is an attractive and suggestive proposal. It clearly emphasizes the importance of self-expression in determining the value of responsible action. The idea that part of the value of responsible agency consists in our ability to express our practical identities, or self, in leading a life seems to be necessary to an adequate account of that value. Moreover, it underlines the limited but significant sense in which our contribution should be understood to fall within a larger narrative, without diminishing the importance of our (limited) contribution.[98]

Section 5.3 – Problems with Fischer's Account

There are a number of difficulties with Fischer's proposal that stem from the connection it makes between self-expression and responsible action. First, generally speaking, it is not clear what the connection between narrative value and the value of moral responsibility is supposed to be. The thought seems to be that artistic creativity and

[98] While Watson (1996) does not make use of the idea of a dramatic narrative, his claim that one sense of responsibility consists in a disclosure of our practical identities can be seen to have strong affinities with Fischer's ideas.

responsible action are both valuable in virtue of the fact that they are instances of 'activity' which discloses a person in some way. This type of disclosure seems possible absent any larger narrative structure. So while it is suggestive to link the value of responsible action to narrativity, I'm not certain the connection is essential or sufficient to cash out the relevant value in question.

Velleman contrasts a situation where someone experiences 10 years of unhappy marriage followed by a divorce and then a happy remarriage, versus a marriage in which there is 10 years of strife followed by a reconciliation between the two individuals and then a lasting happy marriage. Velleman claims that our attitudes towards these two spans of unhappiness (10 years in each case) differ; in the former, the 10 years are viewed as a dead loss, while in the latter, those years are viewed as the necessary building blocks upon which one's future happiness is constructed.

While the impact of a narrative in this case is persuasive I think in showing how dramatic placement shapes our understanding of the nature of well-being (confirming Velleman's point that a simple 'additive' conception of well-being is inadequate), the lesson does not translate for Fischer's purposes. I may only remember one line from a play – 'to be or not to be…' – without much, or any grasp of the place that this sentence plays within the narrative, and yet this seems to me to be a meaningful sentence if uttered by a person in the course of their life. It is certainly true that the meaning may be subject to change or seen as superficial in the absence of further narrative description, but it seems false that the sentence would lack all meaning (and thus all value) should we not place it within a larger context. I need not deny that our lives actually are structured as

dramatic narratives to note that when we sometimes encounter a narrative break, the next act need not be meaningless until we can see how it fits with a dramatic arc.[99]

With respect to the details of Fischer's proposal, I think it is possible to imagine cases of self-expressive action that are at the same time non-morally responsible actions. Such actions, then, would be self-expressive but would lack the particular value associated with morally responsible action. We can express ourselves without thereby being open, in principle, to moral assessment for our attitudes or actions, rendering self-expression insufficient to account for the value of moral responsibility.

Consider the following case:

Mary stubs her foot and in a moment of extreme pain screams 'Ow!'and kicks, accidentally knocking over a prized family possession. While such an action is self-expressive, the experience and outburst accompanying a painful sensation, it is not clear to me that we ought to consider the outburst an instance of morally responsible action. There was nothing impeding the action here or coercing her action; Mary disclosed an aspect of her personality. But, first, it does not seem that Mary discloses all that much of her personality – there is not, I think, all that much relevant *information* regarding Mary's self that is conveyed in this expression of pain.[100] Further, it seems the aspect of the personality that Mary discloses – the sensation of pain and the immediate and visceral reaction – is not sufficient to give us any insight into *her* personality, as a particular,

---

[99] Perhaps Fischer, in holding that a dramatic narrative is essential to rendering our self-expressive activity valuable, has a situation something like that of the characters in Beckett's *Waiting for Godot* in mind. Here I think his case is more plausible, although even in this bleak vision of humanity's predicament, it is not obvious to me that the actions of Estragon, Vladimir and the boy are altogether lacking in value.

[100] This suggests that we might distinguish between a thinner and a thicker conception of self-expression. Since Fischer does not do much unpacking of the concept, I will leave this possible distinction to the side. For my purposes, I think the point is better put as that Fischer's conception of self-expression can't do the work he hopes for it to do.

concrete individual with an unique psychology (rather than as a subject of sensorial experience). The only insight that we gain, I think, is knowledge that pain is generally a form of injury and that it often leads agents to withdraw abruptly from the painful source. But this does not seem informative regarding Mary's agency.

While intuitions may differ here, it is not obvious to me that we would want to hold the individual responsible for acting in the abrupt manner. As a result, we might abjure from holding the individual responsible for having accidentally broken the family heirloom. If this is correct, and it does not seem outlandish that this should be our reaction, then the fact that we can express our present state of mind in a way that does not reflect whatever value is exhibited in morally responsible behaviour should cast doubt on the thought that self-expression is sufficient to answer the question posed at the beginning of the chapter.

If this is the case, then Fischer's case as it stands needs to be supplemented.

Section 5.4 – The Value of Moral Responsibility as Answerability

What can we add to the idea of self-expression that allows us to circumscribe cases of valuable instances of morally responsible action from cases of non-morally responsible action? If we can tease this out, then we might be able to say what it is about responsible acts of agency that are valuable.

I think the distinguishing feature of self-expressive action that is also morally responsible action is when such action expresses something for which we are *answerable*.[101] Consequently, I think we ought to explain the value of responsibility by

[101] A variety of authors have made different use of the idea of answerability – for the use that most closely parallels mine see Smith (2008). I try, however, to show in what way the idea of answerability sheds light on the *value* of morally responsible agency, rather than on more

appealing to the idea of answerability.[102] When we engage in self-expressive action that, in principle, allows for some type of response on the part of an agent, then we engage in behaviour for which we can be held morally responsible. What that response will be will depend on a whole host of factors, including, the other individuals involved as well as the overall narrative structure of our lives. But it is not merely the expression which embodies the value of responsible action. Rather it is expression for which we can request an explanation, a justification, or an acknowledgement of a wrong done, which expresses what is valuable about responsible action. This type of expression always, in a way, opens us up to requests for justification and connects with the fact that when we engage in acts of self-expression, we standardly take them to be justified from our own first-personal perspective.

The value of morally responsibility lies then not in simply the value of self-expressive action, but the value of self-expressive action for which we can intelligibly request justification, explanation, or acknowledgement of a wrong done.[103] The self-responsible person will always consider and potentially respond to this request, but I take it that the value of being responsible is the value of being able to stand in this kind of relationship with other people (the interlocutors of our book), rather than any particular act or practice that the person may engage in. The value of morally responsible agency partly consists in agents being able to access and realize certain important and meaningful forms of relationship both with our past and future selves – this would be the

---

traditional debates regarding the nature of different sorts of responsibility. Compare my use of the idea of answerability in chapter two.

[102] Compare Andrea Westlund's (2003) use of the idea of answerability in explaining the value of autonomy and autonomous action.

[103] By my lights, this helps to explain the value of *morally* responsible action in a way that would not be open to the amoralist. It is an expression, I think, of what P.F. Strawson drew attention to when he placed the reactive attitudes at the centre of a theory of moral responsibility.

'dramatic narrative' that Fischer underlined – as well as with other individuals with which we have dealings. This view captures one aspect of the social dimension of agency.

So, the value of moral responsibility consists in self-expressive action that is open to requests for justification. In other words, it reflects our ability to express ourselves in a way that opens up the book of our lives to critical scrutiny both by our own (later) self and others. Our lives are an 'open-book' in that we are continually writing sentences in this book, but we are not the sole readers, nor do we necessarily have a final say as to what those sentences mean.

As a result, there are two notable differences between my account and Fischer's. First, my account emphasizes that the value of morally responsible action must be *interpersonal* in character. While Fischer seemed to underline the individualistic character (even if he notes that it is holistic in character in implicating an entire narrative structure) of moral responsibility, this ignores, I think, the fact that what is attractive about moral responsibility, in part, stems from the way we are able to interact as morally responsible agents with other agents. We should look to the ability to enter into relationships with others to understand what is attractive about being responsible, rather than merely focusing on the individual and their solitary book.[104] Considering the perspective of the reader of the book, if only implicit, should help us see why we would need to incorporate an interpersonal element into the account. Indeed, should that reader

[104] Compare one of Scanlon's (1986:167) early contractualist formulations of the scope of morality: "On this view [contractualism], moral judgments apply to people considered as possible participants in a system of co-deliberation. Moral praise and blame can thus be rendered inapplicable by abnormalities which make this kind of participation impossible". My claim regarding the value of moral responsibility owes a debt to Scanlon's version of contractualism, specifically the value of 'mutual recognition' developed later in Scanlon (1998: 162).

be absent, we should even look to our own judgements concerning our past action to see how this process might work, for they would manifest the same features.

Second, I take it that the relevant kind of self-expression is one for which we can intelligibly ask for or request justification, explanation or an acknowledgement of a wrong done. This request allows us to mark off certain kinds of cases from other instances of self-expression. Absent the capacities necessary to comply with a request for justification, then we can cordon off self-expression from morally responsible action.

Finally, given that moral responsibility is interpersonal in character, we will not be an authority on the contents of our book. The meaning of our sentences will have to be able to, at least in principle, stand up to the scrutiny any interlocutor might bring. While this obviously rules out our status as omniscient narrator, it does not render our contributions epiphenomenal either. The meaning of our sentence will crucially depend on the sentences themselves and the content of our preliminary judgements and later responses to requests for justification. This interpersonal character or significance, as I termed it above, is what we might view as the social element of responsibility.

Section 5.5 – Objections to the Account of the Value of Answerability

There are several possible objections to this account. First, one may simply doubt the importance of the idea of answerability or 'justifiability' to our lives. In this vein, Russ Shafer-Landau writes: "our need to justify our attitudes and behaviour to others, and to expect from others a like need that is responsive to our demands for justification …strikes me as a deep but contingent fact about most of us" (Shafer-Landau 167). Citing the example of the immoralist and the amoralist as individuals who, it seems, are impervious to our criticisms and our demands, who do not simply 'see' the reason that

morality offers in those cases where they engage in objectionable conduct, Shafer-Landau concludes that justifiability is not an essential feature of rational agency. We might hope that they would feel the force of the criticism of their action if we could only get them to 'see' reason, but because they may not care about us, they would not feel the rational force of this criticism.

In response, it is not clear to me what role these figures are meant to play in the dialectic as presented here. Certainly the immoralist, but I also think even the amoralist, insofar as they take part in the game of giving and asking for reasons, have an interest in being, in principle, able to justify themselves to some other agent. Abjuring from an interest in morality should not eliminate this interest in reasons altogether – morality is simply one normative domain amongst others. And even if, as it is claimed here, the value of moral responsibility lies in the idea of justifiability, this need not be viewed as the kind of moral value that is being rejected by the immoralist and the amoralist. While, again, it is controversial just whether amoralists make moral judgements, and if they do, what the content of the judgement is, the most plausible interpretation of their behaviour it seems is that moral norms such as 'Do not cause unnecessary pain to others' are not viewed as having any weight. A rejection of these first-order moral judgements seems consistent with holding that the activity of so rejecting can be seen to embody something of value. That is, while amoralists need not be seen to endorse the values associated with morality, what I am claiming is that insofar as they give, ask for and receive reasons, they are engaged in an activity that is of (moral) value. In a sense, they are part of a morally valuable enterprise in spite of their best efforts.

From a different perspective, this rejection (or even absence of judgement) can be revealing. As Scanlon notes, "[a] person who is unable to see why the fact that his action would injure me should count against it still holds that this *doesn't* count against it" (Scanlon 1998: 288 – emphasis in original). The point is that even for an amoralist, it is difficult to avoid engaging in the activity that displays the value I noted above.

As a result, it is difficult to see how we might still view ourselves as agents and at the same time not acknowledge the role that the giving and asking of reasons can play in our lives. Whatever the status of such a possibility (abjuring any interest in the giving and asking of reasons and being an agent), it seems to me that we would lose what is distinctively important or valuable about rational agency if we did contemplate this possibility.

So perhaps what Shafer-Landau envisions is that the idea that justifiability is somehow only important to a few or many agents but not all. While the idea of justifying ourselves to others is attractive, it is not a universally shared desire, and we can live a recognizable life as a responsible agent without acknowledging the force of the requirement. This may well be true, but is not strictly speaking an objection to the idea that the value of responsible agency lies in justifiability. Moreover, it seems, what is important is being able to be able to offer reasons for our actions, rather than possessing any actual ability. What is important is being the kind of thing that can request and offer justifications for its actions.

So while I have no direct argument against the thought that we could lead our lives with no interest in justifiability, this does not strike me as a meaningful possibility. It seems that part of what is worthwhile regarding moral responsibility just is the idea that

one could justify or excuse one's action (or acknowledge a wrong done) to some other person. If one truly lacked this fundamental capacity or ability, then we might reasonably question whether or not we have jettisoned what is valuable in regards to rational agency.

Second, Fischer objects to what could be seen to be a version of my proposal offered by Sarah Broadie that sees the import of a person's self-disclosure as residing in laying bare what the agent took to be good or valuable in her situation. He finds that while such a proposal might be attractive, it cannot make sense of what is valuable of instances of akratic or weak-willed action. In those instances where an agent acts against their better judgment, they engage in voluntary self-expressive action but do not pursue what they believe or judge to be good or valuable in the situation. While he notes that one might object that even in the case of weak-willed action one does what one judges to be in some sense good, even this counter-claim[105] cannot make sense of the fact that we can act in a way we find to be completely without worth.

While one may disagree with the details of Fischer's response, granting it, it does not undermine the proposal on offer here that what is valuable about exercises of responsible agency is the thought that agents are open, in principle, to requests for justification, explanation or acknowledgement of a wrong done. This idea of justifiability would not depend on the particular contents of an agent's judgement, nor on the idea that all judgements that an agent makes (and all actions that she performs) are expressions of what she takes to be of value. If in the limiting case identified by Fischer, of the agent who performs an action which she does not find in any degree good or defensible, the request for justification is nonetheless in good order, then the value of justifiability is still operative. If such a request is intelligible, as I think it is, then whatever we can see as

[105] One to which I am sympathetic – see chapter six.

valuable about such an action will lie not ultimately in its self-expression (which, it seems, would be minimal insofar as the action would be one that the agent does not find to be defensible and so presumably does not reflect their evaluative judgement) but rather in the idea that such an action is justifiable.

Lastly, Fischer offers the objection to my account that it is circular. Fischer worries that:

> Ideally…we should be able to specify this value [moral responsibility] without importing the notion of moral responsibility. And yet to require that one's life story be accessible to normative evaluation in the sense of the appropriate application of the reactive attitudes would do precisely this, for moral responsibility just is rational accessibility to the reactive attitudes (Fischer 2006: 117).

Since my appeal to being open to requests for justification is akin to being accessible to normative evaluation, it seems to make use of a disguised form of moral responsibility to explain the value of the thing itself. This would not be a helpful explanation of the significance of morally responsible agency.

In response, if my objections to Fischer's proposal are successful, then we are not in an 'ideal' position with respect to explaining the value of moral responsibility. The success of the objections raised above should, I hope, lay the groundwork for the proposal that I've offered.

More substantively, it is not clear to me that the circularity Fischer identifies is in fact vicious. While it is true that the concept of justification is tied to the concept of responsibility, it seems that we can shed light on the nature of one through reference to the nature of the other. Resorting to what I termed the interpersonal significance of responsible agency can connect these two (already linked) concepts in an illuminating manner. It can circumscribe the boundaries of valuable instances of agency from non-

valuable instances. Moreover, when we consider both the variety and diversity of relationships that self-expression, when tied to the possibility of justification, allows for, as well as the way that it opens us up to being active participants in a moral community, this seems to make sense of why it is that we ought to find morally responsible action valuable. That is, it ought to account for the special significance of morally responsible action. If this account is informative in the way I have just suggested, then the circularity that worries Fischer ought not to be thought to be fatal to the account.

In conclusion, I've argued that we can shed light on the value of responsible action, at least from a compatibilist perspective, by focusing not only on the expressive character of action but also on the fact that such action is open to requests for justification. This allows us to make sense of the connection between the value of expressive action and the value of morally responsible action and also captures the interpersonal character of the value of moral responsibility. While more needs to be said to develop the view to account for the different ways and contexts in which the value manifests itself, as well as its impact on the free will debate, the proposal on offer, I claim, has the resources to explain the significance of morally responsible agency.

Chapter Six – The Rational Relations View, the 1st-Person Perspective and Akrasia

Section 6.1 – The Rational Relations View, Akrasia and Reasons-Internalism

While akratic or weak-willed action is normally considered in debates regarding the nature of practical reason and in the evaluation of various theories regarding the relationship between reason and motivation, it is nonetheless widely accepted that insofar as clear-eyed akratic action is possible the akratic agents who act are responsible for their behaviour.

What's more, if RR is a specification of a conception of responsible agency as RA purports to be, then it seems it is necessary for it (RR) to be able to make sense of practically irrational action. Recall that RR held:

> 'RR: There is a normative connection between attitudes and actions and our capacity for evaluative judgment such that the presence of this connection determines if and only if we are open to moral appraisal and assessment for our attitudes and actions'

RR makes more specific the general conception of rational agency that was outlined in RA. While RR may be preferable to the best alternative account of RA (Frankfurt's view) and may incorporate a viable and attractive conception of responsibility, this may not be sufficient for it to be regarded as our best conception of responsible agency. As R. Jay Wallace has argued "an adequate conception of rational agency must also provide the resources to make sense of such paradigmatically irrational phenomena as *akrasia*, accidie and the like" (Wallace 2006: 49). This is because "practical irrationality and rational agency are two sides of the same coin" (Wallace

2006: 49). RR will have to show how it can make sense of the fact that an agent acting against her best judgement can still be held responsible for her irrational action.[106]

Recent debates concerning akratic action have questioned a long-standing thesis: that akratic action is always irrational. According to what I call the 'classical' view, akratic action, or acting against one's own best or better judgement, is always a paradigm instance of irrational action.[107] Alongside self-deception, accidie and wishful thinking, akratic action was thought to embody those essential features of irrationality in the practical sphere.

Akrasia, however, has been notoriously difficult to explain. While it appears to be an everyday occurrence, it also seems to undermine some of our most deeply held tenets regarding the intelligibility and explanation of action. Accommodating the possibility of the phenomenon has put pressure on what is otherwise viewed as an attractive explanation of the nature of intentional action. Such paradoxical considerations have pushed some to doubt the very possibility of a clear-eyed case of akrasia.[108] RR, however, is decidedly clear in holding that while we are responsible for akratic action, it is nonetheless irrational.

In this chapter I examine two arguments that, from opposing directions, try to show that RR cannot make good sense of akratic action. In the first instance, I examine an argument of Nomy Arpaly's that questions RR's commitment to the standard claim

---

[106] That an account of rational agency entails that we must be able to make sense of the idea of irrational action is perhaps open to question. For doubts regarding that the thought that an account of the conditions of agency requires the possibility of going wrong (or error) see Lavin (2004).
[107] I borrow the term 'classical' from Joseph Raz. See Raz (1999) for an elaboration of the various commitments of the classical view, of which the irrationality of akratic action is only one aspect.
[108] Scepticism regarding the very possibility of akratic action has a long lineage dating at least to Plato. See Hare (1963) for a contemporary version of these ancient doubts.

that all instances of akratic action are irrational. Given the possibility of what she calls 'inverse akrasia', a plausible and realistic moral psychology sensitive to the complex realities of actual agents should countenance the possibility that akratic action can be, at times, a rational response on the part of an agent.[109] Comparatively speaking, according to Arpaly, acting akratically can be much more appropriate (or rational) than acting enkratically or virtuously (by following our best judgement). If RR cannot allow for the rationality of akratic action, it seems, then it cannot make sense of whether or not such agents are properly speaking responsible for their actions in those cases of 'inverse akrasia'. That is, while Arpaly's position seems consistent with RA as outlined above, it poses a challenge to the account of irrationality that seems to naturally flow from RR.

From a different direction, R. Jay Wallace has argued that in order to account for the possibility of akratic action we must posit a basic capacity for control that agents exercise in instances of choice or decision, a position he labels 'volitionism' (Wallace 2006: 58).[110] In order for an agent to be able to act contrary to her best judgement, there must be certain motivational states (e.g. decision and choice) that are "directly subject to our immediate control" (Wallace 1999:58). Were agents to lack such control, it seems, then akratic action would not be possible. Given that RR seems to tie action to the content of our judgements about reasons (or, in what I take to be an equivalent formulation, our judgements about what is good) it is important for RR respond to this challenge. The alternative is leaving unexplained the idea of how RR can even contemplate, never mind account for, the possibility of akratic action. Since RR does not seem to emphasize the importance of 'immediate voluntary control' for ascriptions of

[109] See Arpaly and Schroeder (1999) for the introduction of the term.
[110] See chapter three for an extended discussion and criticism of another form of volitionism.

responsibility, Wallace's proposed solution is not available as a solution for RR. Some other account of the way in which we are responsible for our weak-willed actions must be put forward.

For the purposes of the chapter, some preliminary comment on the nature of akrasia, reasons for action and rationality are called for. First, I understand akrasia to consist in acting against one's own best judgement. This understanding seems to be common ground among those who hold that akrasia is possible and those who hold it is not possible. It is also common ground among those who hold that akrasia is always irrational and those who hold, like Arpaly, that it is sometimes rational.

According to this account, for example, if Mary, after careful deliberation and thought has decided that it would be best, all things considered, to marry Pat, this would constitute Mary's best judgement: the content of her judgement in this situation would be that she should, all things considered, marry Pat. This last fact regarding the content of her judgement remains true even if Mary has actually deliberated imperfectly – perhaps she forgot to consider some of Pat's many lesser qualities and habits when deliberating or did not consider the fact that she previously has had doubts about the value of marriage – nonetheless, at this time, it is still her best judgement that she should marry Pat. It seems that we could attribute to her such a judgement and while we may argue with her that she has not properly thought the proposition through, it would be hard to deny that the content of the judgement would make the most sense of her subsequent behaviour (such as accepting a marriage proposal and making preparations for the upcoming nuptials). If Mary, however, who has not subsequently changed her mind, while at the ceremony facing Pat, finds herself unable to go through with the marriage, then she is acting against

her best judgement. Thus, Mary would be acting akratically. Indeed, even if, after running away from the ceremony she, in the throes of sadness, ends up marrying the first person she meets at the bar where she goes to drown her sorrows, who by a turn of fate happens to be Pat in disguise, she still acts akratically.

This extended example is not meant to definitively supply an answer to the question 'what is akrasia?' but is solely meant to fix the discussion for the present chapter. [111] Thus I accept Davidson's characterization of akrasia or incontinent action as:

> In doing x an agent acts incontinently if and only if: (a) the agent does x intentionally; (b) the agent believes there is an alternative action y open to him; and (c) the agent judges that, all things considered, it would be better to do y than to do x (Davidson 1980: 22)

One important fact of note about akrasia is that it remains action that is done for a reason, albeit the lesser one. As Davidson claims, akratic action remains intentional action, that is, action done for a reason, even if we find it sometimes difficult to render such action fully intelligible from an individual's perspective.

A second fact to note about akratic action concerns the conditions under which we can attribute reasons to an agent for a particular action. According to an influential view of the nature of reasons, what we can call 'reasons-internalism', we can be said to possess a reason only if it can have (some) motivational force for an agent. Having a reason is necessarily connected, on the internal reasons view, to its ability to have motivational force for an agent (or its ability to affect our 'subjective motivational set', as Bernard Williams puts it).[112] The characterisation of the strength of that force is deliberately

---

[111] I do not pretend for this to be a neutral or complete characterization of akratic action. For an interpretation of akrasia that parses it from weak-willed action, see Holton (1999) for an attempt to separate the two. For a different account of the nature of weakness of will see Watson (1977).

[112] See Williams (1981) for the formulation and for the introduction of the internal/external terminology. See Shafer-Landau (2003) for discussion of the distinction.

vague – thus the conditions for the presence of a (weighty or sufficient) reason have been left vague. However, the ultimate presence of a reason is provided by the connection with some element of a particular agent's psychology.

Reasons-internalism contrasts with reasons-externalism. On an external-reasons view, we can be said to possess a reason only if some fact obtains which counts in favour of a certain option or action. That is, a reason can obtain even if it has no connection with, relation to or impact on an individual's psychology or their subjective motivational set. Simply put, a reason may not in actual fact motivate an agent to action, but for all that, it can remain a reason.[113] Thus, there is no relation of dependence, metaphysical or otherwise, for the external-reasons theorist, between the existence of a reason and a person's psychological makeup.

Now the rationality of akrasia or akratic actions, it seems, is a simple affair for the external reasons theorist.[114] If and when an agent conforms to the reasons which apply to her – which, by hypothesis, are just those which favour an option or action she is considering, and are independent of her particular psychological make-up – then her action should be considered to be rational. But conforming to those reasons need make no mention of what the agent believes to be the best in this situation. That is, if Mary decides to not marry Pat at the last minute, and this is in fact what Mary did have reason to do (given Pat's many lesser qualities and her problematic deliberation), even though Mary's best judgement is that she should marry Pat (that is, her judgement remains

[113] It is not clear that the external-reasons theorist need be committed to the stronger claim that a reason may exist that could not motivate a rational agent. Whether we accept stronger characterization, for my purposes, does not affect the purported manner in which external-reasons theorists are committed to the rationality of akrasia.

[114] It should be noted that not all external-reasons theorists (e.g. Scanlon (1998)) accept the rationality of akrasia. I return to this point below.

unchanged), then Mary did the most rational thing to do or performed the most rational action. Her action was rational, if not by Mary's lights at the time, then at least according to those who later evaluate her action according to the actual reasons which applied in her situation.[115]

The question of the irrationality of action is posed most acutely then by those who hold some form of reasons-internalism. Nominally, the irrationality of akrasia stems from an internal difficulty – one judges 'x' to be better or best, that is, that one has more or most reason to do 'x' and yet one does 'y'. The close, systematic, almost conceptual connection we take to hold between our judgements, our attitudes and our action is broken.[116] The agent's action belies the judgement she herself sincerely maintains. The agent flouts just those reasons which she herself acknowledges that she has – an acknowledgement that seems to be a precondition for the presence of a reason on the internal-reasons view. For the recognition of a reason on the internal-reasons view was directly related to its role in our motivational system. Demonstrating the rationality of akrasia is most difficult perhaps on this terrain and it is just what Arpaly proposes.

Finally, the concept of rationality has itself recently been the subject of important attention.[117] For my purposes, we can understand a clear case of irrationality to consist in an inconsistency or incoherence between various attitudes or judgements (what Scanlon calls 'structural irrationality'). Inconsistency or incoherence, and their opposites, would

---

[115] While it is open to a reasons-externalist to say that determinations of rationality are made based upon what an individual believes is rational, even if the existence of reasons is independent of their psychology, I think one motivation for a commitment to reasons-externalism would hold that what is rational would be determined by the bfalance of actual reasons.

[116] For an explanation of the nature of this systematic connection, see Korsgaard (1996b). For discussion concerning where the akratic 'break' takes place see Rorty (1980).

[117] See Kolodny (2005), Broome (1999), and Scanlon (2007) for discussion of the difference senses and normative status of the idea of rationality.

provide the key to understanding what it means to be rational or irrational. For example, if I judge that 'x' is the best option and then also judge that 'x' is not the best option, my attitudes would be inconsistent. Contrast this understanding with the thought that what is rational is provided by the answer to the question 'what do I have most reason to do?'. More generally, this position holds that rationality consists in being (properly) responsive to reasons. So, if I judge 'x' to be the best option and 'x' is in fact not the best option then I am being irrational. For now, note that while Arpaly seems to hold that the rationality of akratic action stems from the fact that it provides the correct answer to the deliberative question just mentioned, even if the RR theorist accepts Scanlon's characterization of irrationality as inconsistency, she will have to respond and make sense of the appropriateness of the phenomenon of 'inverse akrasia'. That is, while some of the debate regarding whether or not we should contemplate the existence or possibility of 'rational akrasia' may be definitional or terminological, I don't think the substance of Arpaly's challenge is exhausted by such disputes regarding the nature of rationality.[118]

Section 6.2 – Inverse Akrasia and the 'Whole Person'

What exactly is Arpaly's challenge to RR?[119]

---

[118] One way to resolve this as a terminological dispute would be by adopting a distinction T.M. Scanlon makes between 'irrationality narrowly construed' and 'the most rational thing to do' (Scanlon 1998: 25-32). According to Scanlon, irrationality narrowly construed deals with those cases of systematic, structural connection, or lack thereof, between our attitudes and our actions. Irrationality occurs just when someone recognizes a reason for an attitude or action but fails to be properly responsive to that reason in their deliberations and action – properly in the sense of what they ought to see fit and do from their perspective. The most rational thing to do, in contrast, is a substantive claim about what an agent has most reason to do in a situation. The conception of what reasons we possess would enjoin just that course of action which would be best to do in a particular situation. It would be the course of action for which this agent and any agent similar in biography, history and circumstance had most reason to perform. While I find the distinction illuminating and congenial, I will try and leave it to the side to address Arpaly's broader concerns with the nature of our best moral psychology.

[119] Alison MacIntyre also challenges the irrationality of akrasia. See MacIntyre (1990). While Arpaly thinks that MacIntyre's arguments are best interpreted as demonstrating the benefits of

Arpaly's case begins with an expansion of the internal-reasons view. Recall that according to that view whether or not we have a reason to perform an action is dependent upon or connected to our subjective motivational set. If there is something in that set (which we can imagine is composed of a person's desires, beliefs, tastes, hopes, etc.) that the agent would acknowledge as giving them a reason for acting then the agent has a reason to perform that action. Now in the simplest of cases, if an agent, say, wants to have a glass of water, and sees some translucent liquid in a glass, then we can safely say that she has a reason to drink the contents of the glass. However, if unbeknownst to her, that is not actually water in the glass but some other translucent liquid which would be harmful to her and would be unsatisfying, then we would normally be inclined to say that the agent actually does not have a reason to drink what's in the glass, even though the fact seems to move the agent and she retains the desire to drink water.

In order to accommodate these cases where agents are ignorant of facts germane to their deliberations and to the presence or absence of a reason for action, internal-reasons theorists have traditionally modified their view to allow for cases where an agent would acknowledge a change in what they see to be a reason if provided the necessary or relevant information. That is, if they were provided the correct facts, our agent would no longer see a reason (or be motivated) to drink what is in the cup; she herself would acknowledge that she did not have a reason to drink the cup's contents, if only by putting it down or spitting out the liquid. What's more, she would be correct, on the internal-reasons view, in so doing, for she would not properly speaking have a reason to drink the contents of the cup.

---

obstinacy and of sometimes being weak-willed, I'll treat the arguments as for the most part complementary and so interchangeable.

Building on this expansion of the internal-reason view, Arpaly argues that if as a result of further deliberation, an agent would come to see that a consideration that was previously ignored, thought irrelevant or considered beside the point, as a reason for action, then this consideration could count as a legitimate reason for action, even though the agent dismissed it (and was correct in so doing) at the time of deliberation. This counterfactual interpretation of the internal reasons view holds: if 'x' is a reason to Φ and as a result of or after a period of deliberation or reflection could be connected to a feature of an agent's subjective motivational set, then that consideration could be a reason for action (just in case they aren't deluded, misled, etc.).[120] An agent could come to discover that she possesses reasons not previously considered to be legitimate candidates through a procedure of rational reflection and deliberation (or through some other less obviously rational process). While I have not sketched a complete account of procedural rationality, it seems that by following whatever such an account amounts to (and it may consist of various deliberative virtues gestured at above as well as qualities such as imagination and empathy) an agent can come to recognize 'x' is a reason in a way it was previously not.

On this view, if an agent is

> not necessarily clear-sighted about what reasons she has on a particular occasion to act in one way rather than another…the practical judgements that she arrives at will express what she believes that she has most reason to do, but might fail to express what she actually has most reason to do or what it would be most rational for her to do (MacIntyre 386).

An agent's cloudy or uncertain understanding of what reasons she in fact possesses, reasons which she herself would acknowledge as having were she to have deliberated properly, is what inhibits her from seeing what is the most rational thing to do.

---

[120] This is in fact Williams' view (1981). See also (Williams 1995).

Given this expanded vision of what reasons an agent can possess at any point in time, Arpaly's case can come into clearer focus. The akratic agent is one who is motivated by a consideration that she herself would acknowledge goes against her best judgement – recall Mary akratically refusing to marry Pat – a best judgment that while presently cloudy she would recognize as proper if she had deliberated properly or otherwise was more procedurally rational in some way. Mary would acknowledge that she should marry Pat – after all, all things considered she's judged this is the best thing to do – but she cannot bring herself to follow her best judgement. Now, it seems that it could be plausible that it is the belief or desire that Pat is a person with many lesser qualities, or some other relevant fact about Pat's character, which is what prevents Mary from following through on her judgement. This belief in Pat's lesser qualities could be what motivates her to run from the ceremony – and flout her best judgement. Thus a reason that Mary herself would recognize to be the best, had her vision not been obscured, is the one that actually moves her to action, even though it is contrary to her professed 'best judgement'.

According to the Arpaly, if Mary was somehow sensitive to considerations that are in fact reasons for action for her, then this sensitivity can "on some occasions outstrip" (MacIntyre 390) her more intellectual ability to see that they are actual reasons; and if just that sensitivity is what moves Mary to action, then her akratic action will not necessarily be irrational.[121] Rather, Mary would have been moved to act by what she had

[121] While this is contrary to the definition of irrationality offered above, for now I want to bracket this concern and accept Arpaly's claim in order to investigate to what degree it can be vindicated.

most reason to do.[122] Further reflection and deliberation will of course allow Mary to see the rationality of her action, given the available alternatives, and vindicate for her the thought that it is actually best to not marry someone like Pat who possesses many lesser qualities – but in the meantime, according to Arpaly, we cannot criticize Mary for any rational failing. She is not irrational, but on the contrary, acted on the reason that was the most rational thing to do in her situation, even if it is contrary to her best judgement.

Of course, Arpaly is not thereby committed to a wholesale recommendation of akratic action; she can, and does, acknowledge it is more often than not irrational. As Arpaly writes: "every agent who acts against her best judgment is, as an agent, less than perfectly rational" (Arpaly 2000: 491). But, on occasion, "there are cases where following the best judgment would make the agent significantly irrational, while acting akratically would make her only trivially so" (Arpaly 2000: 491).[123] Thus I think Arpaly's ambitions are best understood as being modest in scope, to simply allow for the possibility that akratic action could be the most rational thing to do for an imperfectly rational agent when deliberating.

---

[122] Arpaly refers to this process as a kind of "dawning – cases in which people change their minds, sans deliberation, as a result of a long period of exposure to new evidence" (Arpaly 2000: 508).

[123] Depending on how one interprets 'rational' and 'irrational' in the above, Arpaly's claim can be more or less paradoxical. Contrasting a 'manual' account of rationality, which aims to give advice from an agent's perspective, with a theoretical account of rationality which "tells us when people act rationally and when they do not, so that given a God's-eye view of a person's circumstances, beliefs, and motives, one would be able to tell how rational or irrational said person would be in performing a certain action" (Arpaly 2000: 488), Arpaly holds that while it may always be irrational (or bad advice) to say 'act irrationally!' it may be rational from a third-personal theoretical perspective, given an agent's beliefs, desires and circumstances. This theoretical account would also, it seems, come in degrees. Akrasia could be 'rational' in this sense because an agent could be seen, from a third-personal perspective, to act on 'good reasons' (from their point of view) which nonetheless run counter to their own explicit judgement. While this may not be our best theory of rationality, the thought that akrasia can be rational shouldn't depend, I think, on whether or not we disagree with it. The substance of the challenge, as I see it, is not so much to the theory of rationality favored by RR as it is to the priority placed on the first-personal perspective of the theory.

Arpaly also notes that the charge of akrasia is also sometimes held to be irrational because akratic agents seem to render their belief and desire sets (or their entire psychological make-up) less coherent or consistent than they would have otherwise been had they not acted akratically. Here it is thought that an agent's actions are rational just when they express who they truly are or their true self (a RSV). If an agent has a conflicting set of desires or wants, on this view, they ought to satisfy those that better cohere with the agent's other desires or those that are more important to the agent's self. Acting akratically on this view would privilege those aspects of our psychology that are not as important to an agent's self or cohere less well with other elements of our motivational system.

However, for Arpaly, if an agent is actually sensitive to considerations that do in fact better cohere with and better reflect an agent's self, and she acts on those considerations, even though she has judged another course of action to be the best possible course of action, then it seems her akratic action is in this instance is rational. It displays no psychological incoherence or evaluative inconsistency. Rather, it is what best expresses an agent's true, or whole self. Akrasia, while normally a manifestation of our irrational tendencies, can sometimes be a "safety valve" (MacIntyre 399) for the alert, attentive deliberator.

Huck Finn, Arpaly and others are keen to note, acted akratically when he did not turn his friend Jim (an escaped slave) in to the authorities when he had the opportunity.[124] Huck's 'conscience' (what is taken to speak for his best judgement) clearly suggests that he ought to turn Jim in to the authorities, insofar as Jim is a piece of property. Since Jim

[124] See Bennett (1974) for the introduction of the case of Huck Finn to the philosophical literature.

is a piece of property, he ought to be returned to his 'rightful' owner. Given that Huck does not seem to question this odious conception, his conscience is clear in that he must do his 'duty' and return the property to its rightful owner. When Huck leaves to inform others of Jim's escape, however, he finds that he cannot bring himself to turn in his friend. When the moment comes, he lies and does not inform others that Jim is with him, thus allowing Jim to escape to freedom.

However much we may welcome Huck's magnanimous gesture, should we then try and convict Huck of irrationality in addition to his previous error of judging that Jim ought to be returned to his 'owners'? Indeed, it is sometimes claimed that the classical conception is committed to the counter-intuitive conclusion that an agent is more rational when acting on a mistaken judgement (e.g. turning Jim in) than if they were to act akratically. Thus the classical conception encourages us to compound the original error, embodied in the judgement that he ought to turn Jim in, in emphasizing the irrationality of akrasia. It would seem some fetishistic obsession with the importance of an agent's best judgment forced us to call all instances of akrasia irrational rather than any feature of the situation or of an agent's appreciation of that same situation.

Section 6.3 – The Rational Relations View and the First-Person Perspective

In responding to Arpaly's challenge, then, I think we would do well to try and reconstruct some of the considerations that motivated the classical conception in order to see why it has seemed that all instances of akrasia were thought to be paradigms of irrational action.

The classical conception starts from what is an important story about the explanation of action. Joseph Raz writes that "the central type of human action is

intentional action…[which] is action for a reason…[and] reasons are facts in virtue of which those actions are good in some respect and to some degree" (Raz 1999: 23). Intentional action, or action done for a reason, makes sense of our agency because we act in light of various appreciations of ourselves and our world. What does this mean? It means that actions that agents engage in make sense first and foremost to those agents – there is a story to tell from the inside about what made the action the right one for the agent. This story renders the action intelligible from the agent's point of view and provides the material for an explanation of the agent's action.

Now this explanation is not merely a collection of the causal set of events which led to the movement of the agent but rather an explanation, first and foremost, of the facts the agent took to make the action an intelligible object of choice. It is a normative explanation. What made the action an intelligible object of choice was the fact that there were considerations which showed the action to be good in the eyes of the agent. The agent's reasons, then, for which she acted, were some 'good' characteristic (either good in fact and/or good in the eyes of the agent) which made the option they were deliberating about the one to choose. Those reasons may not have required one action or another but they did render the agent's options eligible in her eyes – one's that might be worth pursuing.

The classical approach outlined here emphasizes the priority of the perspective of the agent in rendering an action intelligible: what did the action look like to the agent at the time of action? If it appeared good, that is, if the good-making facts were taken to be reasons, most likely non-conclusive reasons, for an action, then they rendered that action an eligible one for choice for that agent. The agent's intentional action, explained in

terms of her reasons, is intelligible just when we are able to see what 'good' she saw in her situation.

This emphasis on the first person perspective is not a mere piece of "phenomenological biography" (Wallace 2006: 267). For the classical view, the results of first-personal deliberation represent "sound and perfectly generalizable verdicts about what there is reason for anybody to do" (Wallace 2006: 267). As Wallace notes, this gives deliberative judgement a 'priority' that it does not always possess on differing accounts of the nature of deliberation.

Indeed part of Arpaly's program in attacking the irrationality of akrasia was to make room for an account of moral psychology that did not give pride of place to conscious deliberation. She holds that there is an important difference between conscious and unconscious forms of deliberation. Removing a prejudice in favor of conscious, reflective deliberation in order to make room for action that is abnormal would be a welcome development. She claims the following cases are left out of traditional accounts of agency:

> people who appear to act for rational reasons, even for moral reasons, without knowing that they are acting from them or even denying that they are; people whose irrational beliefs are caused by their emotions or desires without the intervention of their own agency; people whose explicitly declared moral beliefs are at odds with the way they act and feel; people who wonder if they are very rational or very foolish; and people who seem to be alienated from parts of themselves that seem to us to be an important part of them (Arpaly 2003: 29).

There is some justice to Arpaly's complaint – cases such as ones of 'inverse akrasia' have not been central to discussions of moral psychology. That said, the mere fact that these cases have not received adequate attention does not guarantee that reflecting on them will demonstrate that an alternate explanatory perspective, one for

example that necessarily allows for the possibility of 'rational akrasia', is needed. So we must rely on the positive considerations raised above in order to see whether or not the priority placed on the first personal deliberative perspective as emphasized by RR is out of place.

For my purposes, I think we ought to note that the results of first-personal deliberation are not to be viewed as just another element in a person's psychology. As I argued in chapter two, such judgements reflect our view as to what constitutes an adequate or sufficient reason in a situation. While deliberation for RR is not necessarily consciously performed, it holds that our judgements regarding the adequacy of reasons have important consequences for other elements of our psychology, such as our desires or other commitments, just because they are reflective of our activity *qua* agents.[125]

But these judgements are not to be viewed as inert objects that have no connection to or are of no consequence for the other elements of an agent's psychological make-up. While it is true that none of us are perfectly rational beings, this does not impugn the fact that there are important rational connections between our various attitudes and between those attitudes and our capacities for evaluative judgement. RR acknowledges this complex set of connections by emphasizing the holistic character of our judgement-sensitive attitudes both in terms of fixing the content of those attitudes but also in noticing the connections those attitudes have with other elements of the agent's psychology.

Thus when we come to a judgement regarding the adequacy of some reason, it is too quick to then compare that judgement to some other element of our psychology and

[125] One way of lessening the sting of Arpaly's charges is to note that insofar as our first personal judgements are not necessarily explicit, we need not solely identify our explicit judgements with the first-personal perspective.

to see if the latter, as it were, is a better reflection of what the agent 'really' thinks, or coheres better with what an agent holds, or is what an agent would come to acknowledge as the proper reason if she had only deliberated in an appropriate, sufficiently imaginative or sensitive fashion. That is, we should not treat each part of an individual's psychology as though it was not a part of one person's overall perspective. Thus if John generally holds that moral reasons take precedence over reasons of taste – if not eating something that he finds not that tasty would constitute a moral offense, or refusing a gift would be disrespectful – then we should not compare John's standing judgement regarding the adequacy of certain reasons with his distaste for roast beef or his dislike for large bits of pottery as though they were two unrelated items. There is a relationship between the two that must be acknowledged when determining 'where John stands'.

While it is clear that Arpaly is keen to emphasize that we must have in view the 'whole person' when determining what best reflects the agent's overall psychological make-up, it is my contention that she does not take the right view of what that 'whole' should look like. There is an important and central role in our conception of the self for an agent's judgements regarding the adequacy of reasons.

In the ideal case, should an agent judge that 'x', a separate attitude to the effect of 'not x' that the agent otherwise holds should be extinguished or disappear as a result of their judgement that 'x'. This is a descriptive claim regarding the background framework that we employ in order to make sense of the various transitions between attitudes that individual agents hold. The importance of the first-personal judgments stem not from their conscious character but from the fact that we take it that, generally, when we judge that 'x', if there are consequences for other attitudes that we have that flow from that

judgement, then these consequences will standardly follow in the wake of the judgement. So it is not a perverse or narrow obsession with the results of first-personal deliberation that suggests we ought to privilege this perspective but a larger explanatory framework in which we make sense of different aspects of agent's agential activity.

This emphasis on the priority of the first-personal perspective can be best appreciated in examples of 'momentous' decisions. Some decisions can be inflection points that have large ramifications for a person's direction in life. Thus, Mary's decision to either leave school or complete her degree could be seen as such an inflection point. The decision would determine her future career, where she would live, interests in life and perhaps personal relationships. As such, it would involve, I think, what most would consider her whole self. If, however, Mary decides to stay in school rather than leave, it is too quick to then compare her decision to stay, on the one hand, with the aspects of her psychology that pushed her to leave, as though they are two static elements of her psychology which can be examined in isolation. Rather, if her judgement is that, on balance, the reasons supporting staying in school are stronger than those that push towards leaving, while she may be objectively mistaken (say, if she has made a factual error or made a fallacious inference), from her own perspective she will have determined what are the relevant considerations at play. Further, in so doing, this judgement of hers will, in the ideal case, have an impact on other judgements that she holds. She will perhaps not view the reasons that pushed towards leaving school as strong enough to justify a decision to abandon her studies. Once she has determined that those reasons are not strong enough then that constitutes her opinion on the worth of this option (and thus helps make sense of her action).

On this view, the various attitudes we have are not merely functions or dispositions to other states or to actions but reflect our judgements as to the worth of various courses of action. However, over and above this feature of our psychology, "we also have *views* as to which factors 'weigh in favor' of various actions. A psychological account that ignored our views and attended only to how we are disposed would be a very incomplete as a *psychological* account of believing something to be a reason" (Scanlon 2006: 727 – emphasis in original). It seems in this instance that in emphasizing the role of emotions that 'bypass' our own agential activity or emphasize the role of 'alienated' desires Arpaly has perhaps assimilated the elements of our psychology to various dispositions of agents to action.

What seems to be lacking from Arpaly's account of the way in which we ought to view which reasons 'best' represent an agent is the fact that that very same agent has views as to the worth of the various courses of action as represented by her attitudes. When those views conflict with other elements of her psychology, the relationship ought to be seen as dynamic. There is a pressure, a normative pressure on those attitudes to alter given the content of the judgment. Prioritizing her judgement as to which course of action is best is not unprincipled or perverse – it rather reflects the agent's activity in a way that is meant to unify and bring together the diverse elements of her 'whole person'.

More generally, the reason that we ought to hold that there is a normative connection between our capacity for evaluative judgement and the various other attitudes that we hold is because

> It would not make sense to attribute all of these states to us if our mental life did not exhibit a certain pattern: in particular, if our beliefs and plans were not normally responsive to our judgments about the reasons for them. To say that this does normally happen, and that the attribution of these attitudes does therefore

> make sense, is just to say that we are rational agents, in a (so far) purely descriptive sense – that is, we are agents who exhibit a certain kind of psychological organization (Scanlon 2006: 727)

What Scanlon is drawing our attention to is a way to explain the behaviour and attitudes of an agent by embedding them within a larger picture of rational agency. If we lacked such a picture of rational agency, one that implicated the kind of psychological organization just mentioned, then according to Scanlon, much of the basis for attributing beliefs and desires to agents would be lost. In order to allow for such everyday attributions, we must posit a background explanatory framework that exhibits a general pattern – one in which our attitudes are standardly sensitive to our judgements regarding reasons. Thus there is no capricious elevating of our first-personal deliberative perspective over other, more neglected or aberrant aspects of a person's psychology at work in this explanatory program. These other elements can appear in the framework – that they do not form the basis of the framework is fine insofar as they exhibit defects or breakdowns which make them inappropriate as a general exemplar for the broader psychological explanation.

One immediate consequence of the classical approach to action explanation is that it necessarily renders akrasia irrational. It does so because the normative presuppositions of the style of explanation outlined above. When we take something to be good in some way, we are normally attracted to that thing, other things being equal, according to how good that thing is. And, again, other things being equal, when rational, we are motivated to do that thing we take to be good just in proportion to what we take its goodness to be. Our reasons and their rational force, what we take to be good, vary in rational, competent agents with our propensity to act. And we explain agent's actions relying upon that co-

relativity of reason and motivation. Agents normally conform to the reasons they themselves acknowledge to be best. They do this, if only, because the better reason is thought to be more rational in the eyes of the agent.

Our *attitudes'* responsiveness to reasons, and to their differences in worth, is what seems to make the akratic an irrational agent. For the akratic does just what her reason-responsive nature indicates should not be done – in her own eyes, this action is less good (there is less reason to do it) than others, and yet it is done. I think this is why we take akratic action to be often difficult to understand, from our own perspective and for the point of view of others, and invoke notions of weakness and temptation when explaining why an agent acted as she did. When we ask why an akratic agent performed the action that she did, she can still offer some reason(s) why – this is after all what makes akratic action an instance of intentional action. But she can only rationalize the action to a certain degree – there seems to be no good, that is, no conclusive, answer to the question: why did you do 'x' and not 'y' when you thought 'y' better?

What is more, for akratics, we should note, as Davidson originally did, that it is two different things to judge that 'x' is best all things considered and to judge that 'x' ought to be done *simpliciter*. Even in the case where an agent does not change her mind, such a difference in judgement can allow for a variety of factors to intervene and prevent the implementation of the original 'all things considered' judgement.

To summarize this long discussion: RR is part of a larger picture of rational agency that emphasizes the necessity of positing a general kind of explanatory connection between our attitudes and our capacity for evaluative judgment. As a consequence this renders the deliverances of our first personal judgement privileged with respect to other

elements of our psychology, insofar as those other elements are thought to be sensitive to that judgement. Further, this renders instances of akratic action irrational.

Arpaly's challenge, however, is not merely ignored. Rather, the motivation for entering this logical space is undermined: we need not posit the possibility of 'rational akrasia' in order to explain the behaviour and attitudes of aberrant agents such as Huck. Nor does RR have the curious consequence of somehow condemning Huck to the thought that he must, on pain of irrationality, follow his 'best judgement'. It can be reasonably suggested that Huck changed his mind – not due to the direction of some arbitrary external source, but from other elements of his psychology – further, note that in fixing the actual content of his judgement, reference to other aspects of an agent's psychology will be incorporated in this process. Part of Huck's attraction as a character, I think, is the fact that he is conflicted. Consequently, there are unappreciated resources, we might imagine, that Huck could have relied upon in order to see for himself that his judgement was mistaken.

It seems in coming to determine just what Huck's 'conscience' is demanding of him that it matters that he has alternate and conflicting feelings with respect to the proposition 'I should turn Jim in'. While the disagreement between Arpaly and me does not turn on the actual details of the character of Huck Finn, even if he has been as vividly painted as a real person, the points made here are general: there are resources within RR to make sense of how we can acknowledge and incorporate the various elements of a person's psychology, including aberrant and other recalcitrant elements, in an explanation of their behaviour. We can provide a picture of the 'whole person', including their conflicting elements, and retain its commitment to the priority of the first-personal

perspective. While RR does not thereby allow for the possibility of 'rational akrasia' and holds that all instances of akratic action are irrational, on my view, there is no real loss in giving this up. What's more, RR can retain the attractive picture of rational agency at its heart.

Section 6.4 – Clear-eyed Akrasia and the Will

Turning to Wallace, his challenge to this picture of rational agency is direct: he holds that it cannot properly account for how *clear-eyed* akrasia is in fact possible. While RR, as well as most other theories, are keen to acknowledge that akratic action is common and perhaps even widespread, it does not possess the theoretical resources to explain how such behaviour is possible. As a result of this inability, Wallace holds that we must posit a more robust form of volitional control as a basic feature of agency in order to explain how it is we can be motivated to act contrary to our best judgement.[126]

Wallace's scepticism stems from his perceptive and sympathetic reconstruction of Scanlon's version of RR. He notes that Scanlon's account of the nature of responsible agency "amounts to a cognitivism about motivation" (Wallace 2006: 271). This form of cognitivism embodies the priority of deliberative judgment described above. However, in the case of akratic action, it seems that RR must trace the akratic's "motivation to act against one's better judgement…to the agent's normative thought that doing so would be good or recommended" (Wallace 2006: 271). The akratic agent, insofar as their action is intentional, would act on a reason that, while not recognized to be best or even better than

[126] While I treat Wallace's argument as dealing primarily with akrasia, I'm not sure it is intuitions associated with akratic thought that is driving the opposition here, or the claim that "[i]f we concede that desires are not subject to a person's voluntary control, then it seems misguided to blame the person merely on account of them" (Wallace 2006: 274). That is, while I can't establish this point, it seems Wallace's overall scepticism regarding RR stems from a distinct conception of responsibility rather than more localized disagreements over the nature and explanation of akratic action.

its competitors, nonetheless would have something to recommend it to the agent. All judgements, for Scanlon and for RR, are of the form that there is something to be said in their favor (however minimal or unimportant it may be).

But this fact, when combined with the role that the priority of deliberative judgement plays, puts pressure on the idea that clear-eyed akratic action is possible. This is because a sincere judgement that 'x' is the thing to do would normally produce the intention to perform this action. Insofar as this intention is not realized in action and the agent acts akratically, Wallace claims, RR must hold that the "normative cognition latent in the wayward desire on which the *akrates* acts must have somehow clouded their grasp of what there is most reason to do" (Wallace 2006: 272 – emphasis in original). This, it seems, is quite close to a denial of the possibility of *clear*-eyed akrasia, insofar as akratic action only takes place under the condition that the agent's grasp of her better judgement is somehow impaired by another element of her psychology. An agent clearly acknowledging that 'x' is a better judgement but opting for 'y' does not seem to be a live option for RR.

Wallace notes that Scanlon marks an important distinction between a judgement concerning the adequacy of reasons and its subsequent effect on an agent's thought and behaviour. What often explains akratic action for Scanlon is the fact that depending on the circumstances, the considerations that led to the formation of our judgement can be more or less vividly represented in an agent's mind. Insofar as a 'wayward desire' is or contributes to a circumstance that renders those considerations less vivid (and perhaps brings others, such as the prospect of pleasure, to the fore), Scanlon has an explanation for why we act akratically and how this is possible.

While this explanation, Wallace contends, is fine as far as it goes, it does not extend far enough in his eyes. He holds that it leaves unexplained many different cases where an agent acts in accordance with their judgement, but while being tempted by competing considerations – cases of strength of will. For example, if I judge that it would be best to not eat ice cream, but am constantly tempted to do so by the fact that my partner always purchases ice cream and keeps it in the house, if I overcome these urges and stay true to my original judgement, then I would be exhibiting an instance of strength of will. In these cases, an agent's attention and thought are pulled away from the considerations that led them to form their original judgement by other wayward desires, desires which are often tantalizingly tempting for an agent. In the face of such temptation however the agent resists and follows through on their original judgement.

To properly account for this phenomenon Wallace thinks that we must posit a more basic form of agency, what he calls "an executive capacity for self-determination in the realm of action" (Wallace 2006: 273), in order to explain what it is the strong-willed agent does when they resist temptation and hold true to their original judgement. A proper account of responsible agency must posit "a capacity to determine what one shall do in ways independent from one's merely given psychological states" (Wallace 2006: 59-60).

Section 6.5 – Can the Rational Relations View account for the Akratic and the Enkratic Agent?

Wallace's objection to RR is thus two-fold. On the one hand, RR assimilates all forms of akratic action to a picture where it is necessarily a cloudy grasp of the original judgement that accounts for acting contrary to our better judgement. This is false insofar

as it leaves unexplained instances of clear-eyed akratic action. While Wallace does not list or offer examples of such action, having a second cigarette (or another beer, or a second desert, or procrastinating during a weekend) all might serve as plausible examples of clear-eyed akratic action. Moreover, strong-willed action is left a mysterious phenomenon. Only positing a more basic capacity for choice can explain the possibility of strong-willed behaviour as well as all instances (not just those that are cloudy) of akratic action.

Consequently, RRs claim to embody an adequate account of responsible agency that is at base non-voluntary is put into question. For such a conception would not be able to make sense of the possibility of acting akratically – more generally, it would not be able to make sense of irrational behaviour in the practical sphere.

RR's response to Wallace's alternative explanation of irrational behaviour in the practical sphere should start from the question: is the picture of agency posited in fact inadequate to account for contrary to better judgement action and strong-willed action in particular? More generally, do these phenomena push us to look for an alternative explanation over and above the resources that are latent within RR's conceptual framework?

RR is able to accommodate the phenomena because the connection it posits between our capacities for evaluative judgement as well as our other attitudes is fundamentally normative in character. Such a connection is not the conceptual or definitional connection that some philosophers held to be between our judgement and

subsequent attitudes.[127] On a normative interpretation, along the lines I develop in chapter two, it is entirely reasonable to hold that we standardly intend to do what we judge to be best in a situation (or better than relevant alternatives) and yet we sometimes do not end up following through on that intention (and acting on a weaker reason that we had previously recognized and dismissed). Normative connections, while perhaps causal (depending on our favored theory of the mind), are not absolute – counter normative behaviour is not as mysterious as it might first appear. Definitional connections cannot contemplate any point at which there can be a break between judgement and attitude.

What's more, helping themselves to the fact that our judgement that 'x' is the thing to do is not equivalent to properly having grasped the depth and significance of such a judgement and its attendant reasons can provide RR with an explanation as to why certain agents act akratically. Their judgements are cloudy and clouded by considerations which they themselves otherwise find attractive or compelling.[128] Indeed, the fine-grained moral psychology outlined in chapter two that distinguishes between seemings, judgings and optings (Scanlon 2002) should provide, I think, the necessary 'moments' to be able to intelligibly see where the akratic break takes places (and whence its source). There is a seeming, which, while not judged to be worth doing overall, jumps the queue as it were in crowding out other considerations when the time to perform an action comes around.

Thus it may seem to John that he ought to have a second drink, even if his all things considered judgement is that given the context – it is a work party he is attending

---

[127] Cf. Hare: "It is a tautology to say that we cannot sincerely assent to a … command addressed to ourselves, and *at the same time* not perform it, if now is the occasion for performing it and it is in our (physical and psychological) power to do so" (Hare 20 - emphasis in original).

[128] This is a kind of strategy that others have adopted to explain the possibility of akratic action. See Tenenbaum (1999) in particular for an historically informed attempt to make use of this kind of distinction.

and he does not always hold back his impolitic views after having drunk too much – the seeming may overwhelm and crowd out these other considerations when the time to decide comes, leading John to blurt out: 'All right, one more it is!', and thus to opt for a drink, against his better judgement.

With respect to the strong-willed individual, we would have recourse to the same style of explanation. While vividly tempted by alternative considerations, does the fact that the strong willed individual does not act akratically call for a complete revision of our theory of agency? It seems not. Rather, what is called for is a psychological investigation of that particular individual and whatever powers of attention, focus and consideration that allowed them to overcome a temptation or interference in their planning and practical life. This, I think, should be sufficient to explain the difference between the akratic agent and the strong-willed agent (as well as to account for the difference between the latter and the 'virtuous' agent).

While Wallace complains that this eliminates the possibility of true 'clear-eyed' akrasia, I think intuitions here are not clear, and we may just as easily be pushed to acknowledging that the true instances of 'clear-eyed' akrasia that Wallace wants to latch on to are more a philosophical bogeyman than a real phenomenon. This is a controversial claim; RR need not offer it. What RR can do is claim that it can explain all the relevant features of all the relevant cases of akratic action in terms of a kind of interference in thought – the interference in this instance being a reflection of a weaker reason (or 'wayward' desire) from another part of our psychological set. Other forms of irrationality can be seen, I think, to be much more controversial. Whether disagreement here should

suggest that we jettison what is otherwise an attractive theory of agency, I hold, is doubtful.

More broadly, RR's account of rational agency can both account for when and why we are responsible for our irrational actions without positing a more primitive form of agency in our capacity for choice or 'executive self-determination'. Indeed, the resources of RR are greater than Wallace's in making sense of our practices of holding agents responsible.

Recall that much of our ordinary thought regarding responsibility attributions do not track the primitive form of agency identified by Wallace or our capacity for self-determination. Much of our ordinary attributions of responsibility track things such as our involuntary responses to circumstances, our emotional reactions to various individuals and developments, our spontaneous outbursts and offerings, what we notice or what occurs to us at certain times, what we do not notice or what doesn't occur to us, and our lapses and forgettings in a number of different contexts. While this is not an exhaustive list, they constitute a strong case for noting that much ordinary practice of responsibility attributions does not presuppose the presence of a capacity for basic choice, but rather a conception of responsibility that can make sense of how we are responsible for non-voluntary manifestations of our agential activity. RR is one version of such an account.

Thus there is a strong presumption against the kind of account that Wallace offers. And given the resources and attractions of RR elsewhere in its ability to make sense of a much larger swath of our intuitions regarding responsibility attributions in other parts of our practical life, as well as its ability to offer what I take to be a reasonable explanation for the possibility of counter-normative behaviour, it seems that we do not have the

necessary challenge required to jettison RR's picture of rational agency. What's more, the theoretical pressure that led Wallace to posit a basic capacity for executive self-determination is released.

To briefly summarize: Against Arpaly I've argued that RR can acknowledge and give proper due to aberrant and transgressive features of our moral psychology while preserving the sense that akratic action is always irrational. Further, contra Wallace, I've claimed that the picture of rational agency that is presupposed in that discussion is sufficient to explain how akratic action is possible without positing a more primitive form of volitional agential activity.

Chapter Seven – Conclusion

Harry Frankfurt has written that the attitude of caring

> is important to us for its own sake, insofar as it is the indispensably foundational activity through which we provide continuity and coherence to our volitional lives. Regardless of whether its objects are appropriate, our caring about things possesses for us an inherent value by virtue of its essential role in making us the distinctive kind of creatures that we are (Frankfurt 1999e, 162-163).

In many ways this comment on the nature and importance of caring reflects the basic commitments of the rational relations view: it is a systematic view regarding the nature of an extended, ongoing self. This view holds that our attitudes – our evaluative judgements as to what is good and what is not – are the ties that, over time, bind together our whole self. They make us the individuals that we are.

But they also provide the basis upon which we are to be thought of as responsible agents. The central role that our judgement-sensitive attitudes play in the constitution of the self provides the raw material for an account of what it is to be responsible for an attitude or an action. By attending to the various aspects of the psychological make-up of a person – what I termed the 'whole person' above – we can come to see that the conditions that make us agents at the same time render us morally responsible agents. The conditions of attributability of attitudes provide at the same time the conditions of moral responsibility.

What this thesis has shown is that not only is the rational relations view the best interpretation of the core connection between attributability and moral responsibility – certainly preferable to its most well developed alternative, offered by Frankfurt – but that it can display a wide variety of theoretical resources to explain and account for a variety of different areas of concern. The rational relations view can help shed light on the nature

of responsibility and its relation to questions of fairness and, to a certain degree, punishment and sanction. It can shed light on the significance of moral responsibility. It can explain why we find it valuable to be morally responsible agents and what role this value plays in our lives. Finally, it can explain how we are responsible for basic features of our agency, including our irrational and exceptional actions. It is an attractive picture that can be usefully extended to these different areas in the philosophies of agency, action and responsibility.

I think the insights of the rational relations view can be further extended. While this theory of the nature of agency and moral responsibility is meant to be, as much as possible, agnostic regarding the best theory of morality, if it is true then it does provide insight and guidance on a number of important and difficult moral questions. For example, a central feature of the rational relations view is that we can be responsible for attitudes and actions which we did not voluntarily choose, because we can nonetheless connect these attitudes and actions to the evaluative and rational capacities of the agent in question. It is this connection which serves as the mark of responsible agency, not conscious, deliberate choice. But a moral theorist's hesitation to ascribe responsibility merely because of the lack of conscious rational choice in the matter should be lessened if the rational relations view is plausible. What it calls for is more detailed and fine-grained theorizing regarding the nature and value of choice as a factor in moral decision making and moral assessment.

Further, a commitment to the rational relations view can recast concerns regarding the 'epsitemic or knowledge requirement' on the attribution of moral responsibility. It is often claimed that there are two distinct conditions on attributions of moral responsibility:

a control condition and a knowledge condition. On the one hand the rational relations view recasts, I think, our best understanding of what the control condition involves. On the other, however, it also forces a re-conceptualization of the knowledge condition. These two conditions, for the rational relations theorist, will not be distinct or unrelated. An agent's judgements will be a direct reflection of her epistemic powers and her use of the information available to her – just as they will be a reflection of her judgement as to what is good or appropriate or right in a situation. The rational relations view thus provides a unified basis to understand what might appear to be distinct conditions of responsibility.

It is not a new thought to hold that the two conditions are not in fact distinct.[129] Rather, what the rational relations view suggests in particular is that the reason these two conditions are not in fact distinct stems from the fact that attributions of responsibility track our judgement-sensitive attitudes. Our judgement-sensitive attitudes express our powers of reason recognition and reflection in both the epistemic and the practical spheres. While I do not argue for the point here, the thought that our judgements regarding a person's epistemic failings concern how they made use of the powers available to them, and the information at hand, shows how the rational relations theorist can offer a unified account of the traditional conditions of moral responsibility.

Further, the rational relations view can help shed light on some of the moral facets of our agency. In particular, I think the account of agency defended above can help

[129] Michael McKenna has written that it "is probably a distortion to think of them [the control and epistemic conditions] as entirely distinct, since how an agent controls her conduct will be in part a function of her epistemic resources" (McKenna 2012: 12).

explain, at least in part, what is *morally* worrisome about weakness of will.[130] Over and above the fact that it may not be advantageous or useful to an agent to lack self-control, there is an intuition that there is a distinct moral failing on display when an agent fails to follow through on what she acknowledges to be her own better judgement. This intuition has been felt difficult to justify when, for example, we consider cases of 'inverse akrasia' where it is felt to be for the best that an agent did not follow through on a morally objectionable attitude. The rational relations view can shed light on this type of failing, however, by shifting attention to the content of the attitude that is expressed by an agent's attitude, especially in light of the agent's other attitudes and her powers of reflection. The wide view of the whole person that is embedded in the rational relations theory can provide the necessary perspective to ground criticism of an agent when she fails to follow through on her best judgement. It can provide the necessary information to ground a direct criticism of the content of that judgement. It can also show how, in the case of a bad agent who fails to follow through on her objectionable best judgement, that agent can be open to criticism. While she might appear to be better than should she have acted on her best judgement, if the source of her inability to act on her own better judgement itself reflects a distinct moral failing – say cowardnness at being discovered as being part of an objectionable project – the rational relations view can explain both the irrationality of such an action as well as why she is nonetheless not to be praised for failing to act on her objectionable judgement. Again, the rational relations view can inform and provide for

[130] See Hill (1991b) for an early and important discussion of the moral dimensions of akratic action. His account, however, is different from the one offered by the rational relations view in important ways, notably in rejecting the requirement that the akratic agent act contrary to their best judgement.

the possibility of more fine-grained analysis of both problems in the theory of responsibility as well as moral theory.

Theories, especially theories regarding the nature of agency and moral responsibility, are not accepted or rejected as a whole. But nor are they simply vindicated in piecemeal fashion. The case made for the rational relations view is a continuing case to see if it can illuminate our experience as agents and ground and justify our best theory of moral responsibility.

Bibliography

Adams, Robert (1985) “Involuntary Sins” *The Philosophical Review* 94(1): 3-31.

Arpaly, Nomy, and Schroeder, Timothy (1999) “Praise, Blame and the Whole Self” *Philosophical Studies* 93: 161-188.

Arpaly, Nomy (2000) “On Acting Rationally Against One's Best Judgment” *Ethics* 110: 488-513.

Arpaly, Nomy (2003) Unprincipled Virtue: An Inquiry into Moral Agency. New York: Oxford University Press.

Bennett, Jonthan (1974) “The Conscience of Huckleberry Finn” *Philosophy* 49: 123-134.

Blum, Lawrence (2004) “Stereotypes And Stereotyping: A Moral Analysis” *Philosophical Papers* 33(3): 251-289.

Borg, Schaich Jana and Sinnott-Armstrong, Walter (2013) “Do Psychopaths Make Moral Judgments?” in The Oxford Handbook of Psychopathy and Law. Kent Kiehl and Walter Sinnott-Armstrong (eds.), Oxford: Oxford University Press, pp.107-130.

Bratman, Michael (1999) “Identification, Decision, and Treating as a Reason” in Faces of Intention: Selected Essays on Intention and Agency. Cambridge, England: Cambridge University Press, pp. 185-206.

Broome, John (1999) “Normative Requirements” *Ratio* 12(4): 398-419.

Calhoun, Cheshire (1989) “Responsibility and Reproach” *Ethics* 99 (2): 389-406.

Davidson, Donald (1980) “How is Weakness of the Will Possible?” in Action and Events. Oxford: Clarendon Press, pp. 21-42.

Davidson, Donald (2004) “Paradoxes of Irrationality” in Problems of Rationality. Oxford: Clarendon Press, 169-187.

Faraci, David and Shoemaker, David (2010) “Insanity, Deep Selves, and Moral Responsibility: The Case of JoJo” *Review of Philosophy and Psychology* 1 (3): 319-332.

Fischer, John Martin (1994) The Metaphysics of Free Will: An Essay on Control. Oxford: Blackwell Publishing.

Fischer, John Martin (2006) “Responsibility and Self-Expression” in My Way: Essays on Moral Responsibility. Oxford: Oxford University Press, 106-123.

Fischer, John Martin (2012) “Responsibility and Autonomy: The Problem of Mission Creep” *Philosophical Issues* 22: 165–184.

Fischer, John Martin and Ravizza, Mark (eds.) (1993) Perspectives on Moral Responsibility. Ithaca: Cornell University Press.

Fischer, John Martin and Ravizza, Mark (1998) Responsibility and Control: A Theory of Moral Responsibility. Cambridge: Cambridge University Press, 1998.

Fischer, John Martin and Tognazzini, Neil (2009) "The Truth about Tracing", *Noûs*, 43 (3): 531–556

Frankfurt, Harry (1988a) "Alternate Possibilities and Moral Responsibility" in The Importance of What We Care About. New York: Cambridge University Press, pp. 1-10.

Frankfurt, Harry (1988b) "Freedom of the Will and the Concept of a Person" in The Importance of What We Care About. New York: Cambridge University Press, pp. 11-25.

Frankfurt, Harry (1988c) "Identification and Externality" in The Importance of What We Care About. New York: Cambridge University Press, pp. 58-68.

Frankfurt, Harry (1988d) "The Importance of What We Care About" in The Importance of What We Care About. New York: Cambridge University Press, pp. 80-94.

Frankfurt, Harry (1988e) "Identification and Wholeheartedness" in The Importance of What We Care About. New York: Cambridge University Press, pp. 159-176.

Frankfurt, Harry (1999a) "Preface" in Necessity, Volition, and Love. Cambridge: Cambridge University Press, pp. ix-xi.

Frankfurt, Harry (1999b) "Faintest Passion" in Necessity, Volition, and Love. Cambridge: Cambridge University Press, pp. 95-107.

Frankfurt, Harry (1999c) "On the Necessity of Ideals" in Necessity, Volition, and Love. Cambridge: Cambridge University Press, pp. 108-116.

Frankfurt, Harry (1999d) "Autonomy, Necessity, and Love" in Necessity, Volition, and Love. Cambridge: Cambridge University Press, pp. 129-141.

Frankfurt, Harry (1999e) "On Caring" in Necessity, Volition, and Love. Cambridge: Cambridge University Press, pp. 155-180.

Frankfurt, Harry (2002) "Reply to Scanlon" in Contours of Agency: Essays in honor of Harry Frankfurt. Sarah Buss and Lee Overton (eds.), Cambridge, MA: MIT Press, pp. 184-188.

Frankfurt, Harry (2004) The Reasons of Love. Princeton: Princeton University Press.

Frankfurt, Harry (2006), Taking Ourselves Seriously and Getting it Right. Stanford: Stanford University Press.

Hare, R.M. (1963), Freedom and Reason. Oxford: Clarendon Press.

Heath, Joseph (1997) "Foundationalism and Practical Reason" Mind, 106 (423): 451-473.

Hill Jr., Thomas E. (1991a) "Social Snobbery and Human Dignity" in Autonomy and Self-Respect. Cambridge: Cambridge University Press, pp. 155-172.

Hill Jr., Thomas E. (1991b) "Weakness of Will and Character" in Autonomy and Self-Respect. Cambridge: Cambridge University Press, pp. 118-137.

Holton, Richard (1999) "Intention and Weakness of Will" *Journal of Philosophy* 96 (5):241-262.

Jaworska, Agnieszka (2007) "Caring and Internality" *Philosophy and Phenomenological Research*, 74: 529–568.

Kolodny, Niko (2005) "Why be Rational?" *Mind* 114 (455): 509-563.

Korsgaard, Christine (1996a) "Two Distinctions in Goodness" in Creating the Kingdom of Ends. Cambridge: Cambridge University Press, pp. 249-274.

Korsgaard, Christine (1996b) "Skepticism about Practical Reason" in Creating the Kingdom of Ends. Cambridge: Cambridge University Press, pp. 311-334.

Kumar, Rahul (1999) "Defending the Moral Moderate: Contractualism and Common Sense" *Philosophy and Public Affairs* 28 (4):275–309.

Kymlicka, Will (1989) Liberalism, Community and Culture. Oxford: Clarendon Press.

Lavin, Douglas, (2004) "Practical Reason and the Possibility of Error" *Ethics* 114(3): 424-457.

Levy, Neil (2005) "The Good, the Bad, and the Blameworthy" *Journal of Ethics & Social Philosophy* 1(2): 2-16.

Levy, Neil (2011) "Expressing Who We Are: Moral Responsibility and Awareness of our Reasons for Action" *Analytic Philosophy* 52 (4): 243–261.

Maclure, Jocelyn and Taylor, Charles (2011) Secularism and Freedom of Conscience. Cambridge: Harvard University Press.

McDowell, John (1998) "Two Sorts of Naturalism" in Mind, Value, and Reality. Cabridge, Mass.: Harvard University Press, pp. 167-197.

McKenna, Michael and Russell, Paul (eds.) (2007) Free Will and Reactive Attitudes: Perspectives on P.F. Strawson's 'Freedom and Resentment'. London: Ashgate Publishers.

McKenna, Michael, (2005) "Where Frankfurt and Strawson Meet" *Midwest Studies in Philosophy* 29: 163-180

Mckenna, Michael (2008) "Putting the Lie on the Control Condition for Moral Responsibility" *Philosophical Studies* 139:29-37.

McKenna, Michael (2012) Conversation and Responsibility. Oxford: Oxford University Press.

McIntyre, A. (1990) "Is Akratic Action Always Irrational?" in Identity, Character, and Morality. O. Flanagan and A. Rorty (eds.), Cambridge, MA: MIT Press, pp. 379-400.

Moran, Richard (2001) Authority and Estrangement: An Essay on Self-Knowledge. Princeton: Princeton University Press.

Nagel, Thomas (1986) The View from Nowhere. Oxford: Oxford University Press.

Piper, Adrian (1985) "Two Conceptions of the Self" *Philosophical Studies* 28(2):173-197.

Raz, Joseph (1990) Practical Reason and Norms. Oxford: Oxford University Press.

Raz, Joseph (1999) Engaging Reason: On the Theory of Value and Action. Oxford: Oxford University Press.

Raz, Joseph (2001) Value, Respect and Attachment. Cambridge: Cambridge University Press.

Rorty, Amelie Oksenberg (1980) "Where Does the Akratic Break Take Place?" *Australasian Journal of Philosophy* 58 (4): 333 – 346.

Ross, Jacob (2012) "Rationality, Normativity and Commitment" in Oxford Studies in Metaethics. Russ Shafer-Landau (ed.). Oxford: Oxford University Press, pp. 138-181.

Scanlon, T. M. (1986) "The Significance of Choice" *Tanner Lectures on Human Values*. Salt Lake City: University of Utah Press.

Scanlon, T. M. (1998) What We Owe to Each Other. Cambridge, MA: Harvard University Press.

Scanlon, T.M. (2002) “Reasons and Passions” in Contours of Agency: Essays in honor of Harry Frankfurt. Sarah Buss and Lee Overton (eds.), Cambridge, MA: MIT Press, pp. 165-183.

Scanlon, T.M. (2006) “Reasons and Decisions” *Philosophy and Phenomenological Research* 72(3): 722-728.

Scanlon, T. M. (2007) “Structural Irrationality” in Common Minds: Essays in Honor of Philip Pettit. Geoffrey Brennan, Robert Goodin, Frank Jackson, and Michael Smith (eds.), Oxford: Oxford University Press, pp.84-103.

Scanlon, T. M. (2013) “Giving Desert its Due” *Philosophical Explorations* 16(2): 1-16.

Shafer-Landau, Russ (2003) Moral Realism: A Defense. Oxford: Oxford University Press.

Sher, George (2009), Who Knew? Responsibility without Awareness. Oxford: Oxford University Press.

Shoemaker, David (2011) “Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility” *Ethics*, 121 (3): 602-632.

Smart, J. J. C. (1961) “Free-will, praise, and blame” *Mind* 70 (279): 291–306.

Smith, Angela (2000) “Identification and Responsibility” in Moral Responsibility and Ontology. T. van den Beld, (ed.), The Netherlands: Kluwer Academic Publishers, pp. 233-246.

Smith, Angela, (2004) “Conflicting Attitudes, Moral Agency, and Conceptions of the Self” *Philosophical Topics* 32(1 & 2): 331-352.

Smith, Angela (2005) “Responsibility for Attitudes: Activity and Passivity in Mental Life” *Ethics* 115 (2): 236-271.

Smith, Angela, (2007) “On Being Responsible and Holding Responsible” *Journal of Ethics* 11 (4): 465-484.

Smith, Angela (2008) “Control, Responsibility and Moral Assessment” *Philosophical Studies*, 138:367–392.

Smith, Angela (2012) “Attributability, Answerability, and Accountability: In Defense of a Unified Account” *Ethics*, 122 (3):575-589.

Smith, Holly (2011) “Non-Tracing Cases of Culpable Ignorance” *Criminal Law and Philosophy* 5:115–146.

Stocker, M. (1979) “Desiring the Bad: An Essay in Moral Psychology” *Journal of Philosophy* 76: 738-753.

Strawson, Galen (1994) “The Impossibility of Moral Responsibility” *Philosophical Studies* 75 (1/2): 5-24.

Strawson, P.F. (1962) “Freedom and Resentment” *Proceedings of the British Academy* 48:1–25.

Taylor, Charles (1976) “Responsibility for Self” in Amelie Oksenberg Rorty (ed.) The Identities of Persons. Los Angeles: University of California Press, pp. 281-299.

Taylor, Charles, (1979) Hegel and Modern Society. Cambridge: Cambridge University Press.

Taylor, Charles (1993) “Explanation and Practical Reason” in The Quality of Life. Martha Nussbaum and Amartya Sen (eds.). Oxford: Clarendon Press, pp. 208-231.

Tenenbaum, Sergio (1999) “The Judgement of a Weak Will” *Philosophy and Phenomenological Research* 59 (4): 875-911.

Vargas, Manuel (2005) “The Revisionist's Guide to Responsibility” *Philosophical Studies* 125 (3):399-429.

Velleman, J. David (1991) “Well Being and Time” *Pacific Philosophical Quarterly* 72 (1): 48-77.

Velleman, J. David (2000) “The Guise of the Good” in The Possibility of Practical Reason. Oxford: Oxford University Press, pp. 99-122.

Wallace, R. J. (1996) Responsibility and the Moral Sentiments. Cambridge: Harvard University Press.

Wallace, R.J. (2006) “Three Conceptions of Rational Agency” in Normativity and the Will: Selected Essays on Moral Psychology and Practical Reason. Oxford: Oxford University Press, pp. 43-62.

Waller, Bruce (2011) Against Moral Responsibility. Cambridge, Mass.: MIT Press.

Watson, Gary (1971) “Free Agency” *Journal of Philosophy* 68:205-220.

Watson, Gary (1987a) “Free Action and Free Will” *Mind* 96 (382): 145-172

Watson, Gary, (1987b) “Responsibility and the Limits of Evil.” in Responsibility, Character, and the Emotions. Ferdinand Schoeman (ed.), New York: Cambridge University Press, pp. 119-148.

Watson, Gary (1996) “Two Faces of Responsibility” *Philosophical Topics* 24 (2):227-248.

Westlund, Andrea (2003) “Selflessness and Responsibility for Self: Is Deference Compatible With Autonomy?" *Philosophical Review* 112 (4): 483-523.

Wiggins, David (1973) “Towards a Reasonable Libertarianism” in Essays on Freedom of Action. Ted Honderich (ed.). London: Routledge and Kegan Paul, pp. 33-61.

Williams, Bernard (1981) “Internal and External Reasons” in Moral Luck. Cambridge: Cambridge University Press, pp. 101-13.

Williams, Bernard (1993) “Moral Incapacity” *Proceedings of the Aristotelian Society New Series*, 93: 59-70.

Williams, Bernard (1995) “Internal Reasons and the Obscurity of Blame” in Making Sense of Humanity. Cambridge: Cambridge University Press, pp. 35–45.

Williams, Bernard (2002) Truth and Truthfulness: An Essay in Genealogy. Princeton: Princeton University Press.

Wolf, Susan (1990) Freedom within Reason. New York: Oxford University Press.

Wolf, Susan (2003) “Sanity and the Metaphysics of Responsibility” in Free Will. Gary Watson (ed.). New York: Oxford University Press, pp. 372-387.