Moral Responsibility and Quality of Will

by

Andrew Khoury

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved September 2011 by the
Graduate Supervisory Committee:

Peter French, Chair
Cheshire Calhoun
Douglas Portmore

ARIZONA STATE UNIVERSITY

December 2011

ABSTRACT

This dissertation puts forth an account of moral responsibility. The central claim defended is that an agent's responsibility supervenes on the agent's mental states at the time of the action. I call the mental states that determine responsibility the agent's quality of will (QOW). QOW is taken to concern the agent's action, understood from an internal perspective, along with the agent's motivations, her actual beliefs about the action, and the beliefs she ought to have had about the action. This approach to responsibility has a number of surprising implications. First, blameworthiness can come apart from wrongness, and praiseworthiness from rightness. This is because responsibility is an internal notion and rightness and wrongness are external notions. Furthermore, agents can only be responsible for their QOW. It follows that agents cannot be responsible for the consequences of their actions. I further argue that one's QOW is determined by what one cares about. And the fact that we react to the QOW of others with morally reactive emotions, such as resentment and gratitude, shows that we care about QOW. The reactive attitudes can therefore be understood as ways in which we care about what others care about. Responsibility can be assessed by comparing one's actual QOW to the QOW one ought to have had.

ACKNOWLEDGMENTS

For helpful comments, inspiring discussion, and much needed guidance, I thank the members of my committee (both past and present): Cheshire Calhoun, Peter French, Douglas Portmore, and Margaret Walker. For encouraging the curiosity that eventually led to this undertaking, I thank my parents, Chris and Linda. And for keeping me sane during this project, I thank my fiancée, Caitlin.

TABLE OF CONTENTS

Page

PREFACE

Rather than writing a traditional dissertation that begins with a literature review and steadily progresses to a final culminating conclusion I have chosen to tackle the unwieldy topic of moral responsibility with the tactics of guerilla warfare. Responsibility is well suited to this approach for it is a large and messy beast. A systematic and comprehensive treatment of the topic is, in my opinion, bound to distort and oversimplify. What follows is a collection of essays that attempt to make progress on this topic by taking stabs at it from a variety of angles. To stretch the warfare analogy a bit further (too far?), the weaponry employed throughout is the idea that rather than being about *free will* responsibility is about *quality of will*.

Using a different analogy, the picture of responsibility that I have attempted to develop owes much more to the impressionism of Monet than to the realism of Vermeer. Viewed close up there are certainly questions left unanswered and details to be filled in, but I think that when we step back this account better captures most fundamentally what responsibility is about. There is, of course, much more work to be done, but I hope to have laid the foundation upon which more fruitful research can build.

INTRODUCTION

This dissertation is about moral responsibility. This is not a new topic. People have pondered the nature of responsibility for well over two thousand years. What is new, or at least newer, is the approach to the topic taken here. Theorizing about the nature of responsibility has been primarily guided by concerns about the truth and implications of causal determinism. Determinism can be described as the thesis that a complete description of the universe at a time together with the laws of nature entails every future truth. Many have wondered whether free will is compatible with the world being determined in this way. And it is generally supposed that responsibility requires free will.

Theorists have often begun their theorizing about responsibility with a conviction about the compatibility of free will with determinism. The theories that arise, then, are tailor made to buttress this conviction. The methodological assumption shared by both sides of the debate is that the compatibility or incompatibility of responsibility with determinism is a desideratum of the theory. Given this, it is no wonder that this debate seems to inevitably lead to, what John Martin Fischer has called, a dialectical stalemate.

This is not how responsibility is treated here. Though this is a dissertation about responsibility this is not a dissertation about free will. The reader will notice that, aside from a few footnotes and a remark here and there, the compatibility question is rarely mentioned. (Though in the interest of full disclosure I should note that my sympathies lie with the compatibilist). This is because of my belief

that an answer to the compatibility question should not be an antecedent desideratum of the account. For before we can begin to think about that question, we must know what responsibility is about. It is on this front that I think many theorists have gone wrong and on which I hope to make some progress.

The common theme running through this collection of essays is that responsibility is about *quality of will*. This is, I believe, the lesson that should be taken from, perhaps, the two most influential philosophers working on responsibility in the 20th century: Strawson and Frankfurt. Both Strawson's "Freedom and Resentment" and Frankfurt's "Alternate Possibilities and Moral Responsibility" invite us to question, in different ways, the conviction that responsibility is about free will. Once we are free of this preoccupation a new picture of responsibility emerges, one that is embedded within a social fabric and whose richness accurately reflects the complexity of our moral lives. The essays that follow begin to explore and develop this way of thinking about responsibility.

The first chapter begins with a discussion of Strawson and sets out the methodology employed in this collection. The second chapter begins to develop my conception of quality of will and shows the very close connection between Strawson and Frankfurt. The third chapter argues that the only appropriate target of responsibility is quality of will. Chapter 4 examines the relationship between blameworthiness and wrongness. I argue that rather than requiring wrongness in fact, blameworthiness requires that one ought to have believed one's action to be wrong. Chapters 5 and 6 defend quality of will accounts against what are often thought to be the most challenging objections. Chapter 5 address the "tracing

cases” by applying some of the points developed in the third chapter. Chapter 6 addresses objections involving brain manipulation by distinguishing between responsibility at the time of action and responsibility at some time after. Chapter 7 applies the distinction made in the previous chapter to issues surrounding the notion of collective responsibility. The collection ends with Chapter 8 which seeks to understand holding responsible on the model of speech acts. In many ways, it is a return to the methodological questions that were taken up in the first chapter.

CHAPTER 1

BEING RESPONSIBLE AND HOLDING RESPONSIBLE

"Only by attending to this range of attitudes can we recover from the facts as we know them a sense of what we mean, i.e. of all we mean, when, speaking the language of morals, we speak of desert, responsibility, guilt, condemnation, and justice."

-P.F. Strawson, "Freedom and Resentment"[1]

Readers of P.F. Strawson's justly influential "Freedom and Resentment" have taken the above passage, and others like it, to be expressing this idea:

Being responsible is to be understood in terms of holding responsible.

I here want to distinguish between three different ways that we can understand this idea. First, we can understand this to be making a strong conceptual claim about the conditions of responsibility. Call this *the strong conceptual reading* (SCR):

(SCR) To be responsible just is to be a prone target of the reactive attitudes.

According to (SCR) to be responsible is simply the propensity to be treated as such. On this view there are no independent conditions that warrant such treatment. On this view, we might say, being responsible is conceptually exhausted by holding responsible. This view can be distinguished from a weaker

[1] Strawson (1962), reprinted in Watson (1982, p. 78). All references to "Freedom and Resentment" will be from Watson (1982).

conceptual relation between being responsible and holding responsible. Call this *the weak conceptual reading* (WCR):

> (WCR) To be responsible just is to be an appropriate target of the reactive attitudes.

According to this view being responsible is a matter of being appropriately held responsible. WCR, like SCR, links being responsible conceptually with holding responsible. But unlike SCR, WCR does not hold that being responsible is fully explained by mere holding responsible. Rather, there must also be reference to the appropriateness of holding responsible. On this view there are conditions of being responsible but they do make essential reference to holding responsible. Finally, we can understand the claim as making an epistemic rather than conceptual point. Call this *the epistemic reading* (ER):

> (ER) Inquiry into the nature of responsibility should proceed via the stance of holding responsible.

(ER) makes a methodological rather than conceptual claim about responsibility. It says that theorizing about responsibility should begin by reflecting on what it is like to hold responsible. And it is consistent with both (SCR) and (WCR) though it entails neither. In what follows I discuss the implications of these three interpretations of the way in which holding responsible is prior to being responsible. I argue that while many have attributed to Strawson the conceptual readings, his real insight is (ER).

Strawson's Account

"Freedom and Resentment" begins theorizing about moral responsibility from an angle that had not been traditionally taken. For Strawson, inquiry into the nature of responsibility cannot occur in a social vacuum. Rather, it must be placed within the context of our web of social practices and relationships. At the heart of these practices and relationships lie the reactive attitudes. He begins by focusing on, what he calls, the participant reactive attitudes. These are the emotional reactions we feel in response to the attitudes of another directed towards us in action. For example, one's reaction to being pushed by another will be very different depending on the attitudes expressed by that act of pushing. It is one thing if the pushing expressed indifference towards one's well being, and quite another if it expressed concern about the oncoming bus in one's path. These are to be distinguished from mere moral condemnation or approval, in that the latter but not the former, lend themselves to "a certain detachment from the actions or agents which are their objects" (Strawson, 1962, p. 62). The reactive attitudes are reactions to the intentions and actions of others, and they are significant because of the great importance we place on the actions and intentions of others. In general, we demand a certain amount of good will in others, and that amount depends on the relationship we stand to that person. When we recognize that an individual has met or exceeded this demand for good will, we will feel things like gratitude, and when we recognize that that demand has not been met, we may feel resentment.

Strawson identifies a class of reactive attitudes that are reactions to the quality of will of another that is expressed at a third party. These are vicarious analogues to the participant attitudes. The third person version of resentment is indignation, and it is this third person nature that gives these attitudes the title of moral. We feel indignation when we believe that an agent has not met the demand for goodwill in his dealings with another. This is not to say that the moral reactions are essentially vicarious, or third person in nature. One can feel indignation when it is oneself that is wronged, and so one can feel that she herself has been the victim of a moral transgression. The point is that they are essentially capable of being vicarious, not that they essentially are vicarious. It is the fact that these attitudes generalize that makes them moral. Like the personal reactive attitudes, which are an expression of a demand that others treat us with good will, these attitudes represent a generalized form of that demand. Strawson then identifies a third class of reactive attitudes, the self-reactive attitudes. These are reactions to one's own quality of will such as guilt and shame. He argues that the three classes are "humanly connected" (Strawson, 1962, p. 72) in that they come as a package deal for normal adult human beings. For Strawson, the stance of holding responsible, which consists in the reactive attitudes, has special relevance to being responsible. Holding responsible is, in a particular way, prior to being responsible. But which way, exactly, is a matter of contention.

Interpreting Strawson to be Making the Strong Conceptual Claim

According to (SCR) to be responsible is simply to be held responsible. Gary Watson (1987) seems to interpret Strawson as holding (SCR):

> As his title suggests, Strawson's focus is on such attitudes and responses as gratitude and resentment, indignation, approbation, guilt, shame, (some kinds of) pride, hurt feeling, (asking and giving) forgiveness, and (some kinds of) love. All traditional theories of moral responsibility acknowledge connections between these attitudes and holding one another responsible. What is original to Strawson is the way in which they are linked. Whereas traditional views have taken these attitudes to be secondary to seeing others as responsible, to be practical corollaries or emotional side effects of some *independently comprehensible belief* in responsibility, Strawson's radical claim is that these "reactive attitudes" (as he calls them) are constitutive of moral responsibility; to regard oneself or another as responsible just is the proneness to react to them in these kinds of ways under certain conditions. There is no more basic belief which provides the justification or rationale for these reactions (Watson, 1987, pp. 256-257, italics mine).

And later:

> In Strawson's view, there is no such independent notion of responsibility that explains the propriety of the reactive attitudes. The explanatory priority is the other way around: It is not that we hold people responsible because they *are* responsible; rather, the idea (*our* idea) that we are responsible is to be understood by the practice, which itself is not a matter of holding some propositions to be true, but of expressing our concerns and demands about our treatment of one another (258)…These nonpropositional responses are constitutive of the practice of holding responsible" (Watson, 1987, p. 261).

Watson interprets Strawson to be a noncognitivist about our practice of holding responsible. Insofar as the reactive attitudes have no propositional content, they permit no justification. As such, our search for some "independently comprehensible belief in responsibility" is misguided. The best we can do, according to this interpretation, is simply to observe the practice to see what contexts elicit these attitudes. The susceptibility to the reactive attitudes fully explains being responsible. No judgment that one is responsible grounds, or

explains, or justifies the reactive attitudes. Rather, having the reactive attitude explains the judgment that one is responsible.[2]

On the face of it, (SCR) is implausible. The degree of relativism and moral infallibilism that it implies will be too much for most of us. Furthermore, I don't believe that this is an appropriate interpretation of Strawson. This is because the reactive attitudes, by Strawson's own lights, do have propositional content. They are about the moral quality of will expressed by an agent in her action.[3]

Interpreting Strawson to be Making the Weak Conceptual Claim

R. Jay Wallace (1994) develops a subtle and insightful Strawsonian account in his *Responsibility and the Moral Sentiments*. "If we wish to make sense of the idea that there are facts about what it is to be a responsible agent, it is best not to picture such facts as conceptually prior to and independent of our practice of holding people responsible" (Wallace, 1994, p. 1). Rather, he claims that our understanding of responsibility should be given *a normative interpretation*. According to this the "conditions of responsibility are to be construed as conditions that make it fair to adopt the stance of holding people responsible"

[2] R. Jay Wallace interprets Watson's essay in this way: "Gary Watson has taken Strawson to be making the 'radical claim' that the reactive attitudes 'are constitutive of moral responsibility; to regard oneself or another as responsible just is the proneness to react to them in these kinds of ways under certain conditions'…Construed along these lines, Strawson's own approach appears to have a markedly noncognitivist character" (1994, p. 74), which, Wallace thinks, "makes the claim hard to accept" (1994, p. 11).

[3] This point will be elaborated in greater detail later.

(Wallace, 1994, p. 15). Holding responsible, in turn, involves either an actual susceptibility to the reactive attitudes or belief that such attitudes are warranted.

Unlike Watson, Wallace takes Strawsonian theory to be an essentially cognitivist enterprise. This is because he sees the reactive attitudes as sharing a common propositional object: the belief that the agent has violated a moral expectation that one accepts. The reactive attitudes are fair, according to Wallace, so long as that expectation has actually been violated.

While Wallace's approach escapes the worries associated with noncognitivism, Angela Smith (2007) raises doubts about any account that gives conceptual priority to holding responsible over being responsible. She criticizes Wallace's theory because she sees the conditions that warrant holding responsible, what she calls active blame, as requiring more than what is intuitively thought of as being responsible. That is, she points to cases in which it is intuitive that the agent is responsible, and yet it would be inappropriate to actively blame her.[4] For example, two people may both judge an agent to be blameworthy and yet it may be appropriate for one but not the other to actively blame the agent due to the relation she stands to the agent (perhaps one is the victim of the harm while the other is merely a bystander). Thus, it is a mistake to construe the conditions of being responsible as equivalent to the conditions that make active blame appropriate.

[4] For Smith, active blame only requires the occurrence of inner attitudes such as resentment. See next footnote.

Wallace could respond in a few different ways. First, he could stress that his conception of holding responsible entails no overt behavior. Rather, to hold responsible involves an actual susceptibility to the reactive attitudes or the belief that such a reaction is warranted. The cases that Smith points to rely on the inappropriateness of expressing one's reactive attitudes but they don't seem to show that the presence of the reactive attitude, much less the belief that such an attitude would be warranted are themselves inappropriate.[5]

Secondly, Wallace could admit that she has successfully brought up cases in which though it is intuitive that the agent is blameworthy it is also intuitive that blame would be inappropriate. But he could argue that this doesn't show his normative interpretation to be misguided because inappropriateness doesn't entail unfairness. That is, though it may be inappropriate in a case to engage in overt blaming behavior, such behavior would not be *unfair*.

What we should notice though, is that the issue is whether the class of cases in which it is fair to hold an agent responsible is coextensive with the class of cases in which an agent is responsible. If Wallace construes holding responsible to be an active and overt behavior then though the normative

[5] Interestingly, Smith also allows that active blame, in her sense, requires no overt behavior: "But on my view one can 'actively blame' a person simply by feeling resentment, indignation, or anger toward her, without ever expressing these emotions in any way" (Smith, 2007, p. 477). As such it is hard for me to imagine a case in which an agent is blameworthy and yet it would be inappropriate to feel resentment. And it is even harder to imagine a case that falsifies Wallace's weaker disjunctive conception of holding responsible. For that would be a case in which though the agent is blameworthy it would be inappropriate to feel resentment or to believe resentment is warranted. But it seems that to say that an agent is blameworthy simply is to say that blame is warranted.

interpretation is interesting and original, it seems vulnerable to the cases Smith raises. If, however, Wallace's account is to be intuitively adequate then he must have a weaker conception of holding responsible which begins to undermine interest in the normative interpretation. For as Smith puts it, "Presumably, what would be 'fair' would be to judge people to be culpable when they are in fact culpable, and that would be determined by asking whether they have transgressed any moral norms or requirements" (2007, p. 472). Put in a different way, Wallace's normative interpretation (what has here been characterized as WCR) has bite only if fairly holding responsible is a narrower class than what we intuitively think of as being responsible. But this will seem to be an implausible account of being responsible for the reasons that Smith raises. However if fairly holding responsible is understood in a broader way that matches up with the intuitive idea of being responsible, then it is hard to see what kind of priority holding responsible is supposed to have. For rather than thinking an agent responsible because it would be fair to hold her responsible, it would seem to be fair to hold an agent responsible *because* she is responsible. Given these considerations, we should move towards the idea that holding responsible has epistemic, but not necessarily conceptual, priority over being responsible.[6]

[6] Wallace's own account works best along these lines, though given the emphasis he places on the normative interpretation he does seem to be conceiving the priority to be conceptual.

Interpreting Strawson to be Making the Epistemic Claim

Michael McKenna (2005) offers a different reading of Strawson. He (tentatively) believes that Strawson should be read as denying both of the conceptual interpretations of the priority of holding responsible over being responsible:

> So, by the light of Strawson's own essay, theorizing about moral responsibility begins by focusing first on the attitudes of agents-on the persons who are held morally responsible, not on those who are holding them morally responsible. For Strawson, the morally reactive attitudes are responses to the quality of will expressed in a person's conduct. Since the reactive attitudes, as well as the attendant practices and expectations, constitute what it is to hold agents morally responsible, holding morally responsible is to be understood on Strawson's view as tailored to the moral quality of will indicated in the activities of morally responsible agents. Hence, the explanatory relation between holding and being morally responsible is, at best, mutually supporting. And there is good reason to think that, on Strawson's own view, if one direction has a place of privilege, it is being morally responsible; being morally responsible explains holding morally responsible"(McKenna, 2005, pp. 171-172).

For McKenna, when Strawson says "Only by attending to this range of attitudes…" (1962, p. 78), the range of attitudes that he is referring to includes the attitudes of those *held* responsible in addition to the attitudes of those holding responsible. The reactive attitudes are reactions to the attitudes of the agents who are the objects of evaluation; specifically the degree of ill or good will revealed by their action. So on McKenna's reading, agents are responsible insofar as they act from good or ill will, and the reactive attitudes that constitute holding responsible are a response to the agent's quality of will.

I think that McKenna has it right. There is ample evidence that suggests that Strawson took the priority of holding responsible over being responsible to be epistemic rather than conceptual. In the opening quote of this chapter I take Strawson to be making a methodological claim; that by focusing on the reactive

attitudes we will be in a better position to understand what responsibility is about. "The central common place 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, p. 62). The reactive attitudes are intimately linked to beliefs about the quality of an agent's will. This suggests that the noncognitivist reading of Strawson, favored by Watson, is not accurate. Contra Watson, the reactive attitudes do have propositional content which concern the attitudes and intentions of the agent. "What I have called the participant reactive attitudes are essentially natural human reactions to the good or ill will or indifference of others towards us, as displayed in their attitudes and actions" (Strawson, 1962, p. 67). One might read Strawson here, as McKenna does, as stating that an agent is morally responsible for some action insofar as that action stemmed from good or ill will or indifference. And because the reactive attitudes are reactions to the quality of will of others (or ourselves), and being morally responsible is to be understood in terms of the quality of will of the agent, then being responsible is conceptually prior to holding responsible.

Strawsonian Methodology

On this reading Strawson has not presented a theory of responsibility so much as he has set up an account of how theorizing about responsibility ought to go. In the comfort of our armchair we must not lose sight of what responsibility is about.

We should begin by vividly reflecting on what it is actually like to hold responsible. And notice that we don't, typically, hold agents *merely responsible*. Rather we blame and praise them, we feel resentment and gratitude. This suggests a different kind of conceptual priority. Rather than thinking that holding responsible is conceptually prior to being responsible, we should think of blameworthiness and praiseworthiness as conceptually prior to mere responsibility. I think that this is Strawson's most original insight.

Many theories of responsibility begin by surveying intuitions about the responsibility of agents doing rather mundane things. For example, Fischer and Ravizza ask for our intuitions about whether an agent is responsible for making the train go to Syracuse when nothing of moral importance is at stake.[7] This, I think, can lead to a distorted picture of responsibility. There is a sense in which the question is baffling. Am I morally responsible for raising my right hand in the privacy of my home? It is hard to know how to make sense of this question. For Fischer and Ravizza the question amounts to whether the agent exercised a particular type of control. But I am left wondering whether this is really what responsibility is about.

[7] "Ralph is the driver of a train whose brakes have failed. We suppose, for the sake of the example, that Ralph has been kidnapped and required against his will to drive the train. The train is hurtling down the tracks toward a fork in the tracks. Ralph knows that, although he can cause the train to take the right fork or the left fork, he cannot stop the train. Ralph also knows that both forks lead to Syracuse. When Ralph turns the train onto the left fork, he can be held morally responsible for the consequence, that the train takes the left fork (rather than the right fork). But it just seems obvious that Ralph is not morally responsible for the consequence, that the train ends up in Syracuse, given the fact that Ralph is not morally responsible for the fact that he is on this stretch of track in the first place" (Fischer and Ravizza, 1998, p. 94).

Alternatively we can begin our inquiry by focusing, not on these cases of mere responsibility, but on cases that engage our emotions, cases that elicit strong feelings of resentment on the one hand, and gratitude on the other. That is, we should focus on cases in which we take the agent to be blameworthy or praiseworthy. When I focus on my feelings about the callous murderer what I find is that my resentment consists, in part, in representing the murderer to be a certain way. I am representing him to have cared about the wrong things and to have not cared enough about the right things. I am representing his feelings and attitudes expressed in his actions to be despicable. I am representing his quality of will to be a certain way. When I reflect on the feelings I have about the bystander who threw herself in front of the gunman in order to save another I am struck by the selfless concern that motivated her. Here too I am representing her quality of will to be a certain way, though in this case exceptional. Strawson's insight is that when we pay closer attention to what we are doing when we hold responsible we can see what responsibility is about. It is about quality of will.

Smith, I think rightly, has pointed out that holding responsible in an active and overt way requires more justification than that the agent is in fact responsible. But she takes this to suggest "that these concerns about the fairness or appropriateness of blaming responses should not themselves play a role in our theorizing about the conditions under which a person can be said to be responsible or culpable for her actions and attitudes" (Smith, 2007, p. 483). But granting that the conditions that justify active blame require more than blameworthiness, there is still reason to look towards our practice of holding responsible in order to

understand being responsible. This is because when we are engaged in our most active forms of holding responsible the representation of the agent as responsible is made most vividly. It is in such circumstances that we can best see what responsibility is about. Insofar as the reactive attitudes have propositional content, then the conditions of being responsible simply are the truth conditions of those propositions. But it is from the stance of actively holding responsible that we are in the best epistemic position to see what that content is.

Strawson saw that our practice of holding responsible is emotionally charged. And when we look to these emotions what we find is a representation of the quality of will of the agent. Resentment involves the representation of a criticizable quality of will and gratitude involves the representation of an exceptional quality of will. The reactive attitudes are warranted to the extent that their constitutive representations are accurate. To be blameworthy for an action simply is to have acted with a criticizable quality of will and to be praiseworthy for an action simply is to have acted with an exceptional quality of will. Since the reactive attitudes represent blameworthiness and praiseworthiness these notions are conceptually prior to responsibility itself. The notion of moral responsibility is parasitic upon these concepts which are constituents of the reactive attitudes. To say that an agent is morally responsible for an action is to say that it is the sort of thing that she could be blameworthy or praiseworthy for were there something riding on it. This is the reverse of traditional thinking about responsibility. According to this tradition responsibility, conceived as an agential capacity, is most fundamental. To see an agent as blameworthy is to first see her as

responsible and then to see her as having done something wrong. On this view it is easy to see responsibility as a metaphysical issue and to be led into the "panicky metaphysics" of which Strawson was so wary. But when we give credence to our humanity we find, not that responsibility has metaphysical presuppositions, but that it about the degree of good will we express to one another.

CHAPTER 2

MORAL RESPONSIBILITY AND QUALITY OF WILL

Many theories of moral responsibility give the notion of *quality of will* a central role. Indeed, elements of this idea are near ubiquitous in moral theorizing. Here I develop, in broad strokes, a quality of will based theory of responsibility. This theory holds that moral responsibility for an act is completely determined by the quality of will with which an agent acts. I'll first give a rough sketch of this notion and show the role it has played in the works of some influential philosophers. I'll then begin to articulate a more careful conception of quality of will. An agent's will is, on this account, an action understood internally (what might be better described as a willing rather than the will). The quality of an agent's will is determined by the motivations with which an agent acts. An agent's motivations are determined by the reasons in virtue of which she was actually moved to act. The reasons we do and do not respond to reveal and are determined by what we care about. Our reactions to the quality of will of others, exemplified by the reactive attitudes, shows that we care about quality of will. The reactive attitudes are therefore second-order cares that are aimed at quality of will. This shows that quality of will is what responsibility is about.

1. Historical Conceptions

It is extremely plausible that an agent's responsibility for an act depends, in some way, on her attitudes and beliefs at the time of action:

> Persons interpret each other's movements as manifestations of intention and

> choices, and these subjective factors are often more important to their social relations than the movements by which they are manifested or their effects. If one person hits another, the person struck does not think of the other as *just* a cause of pain to him; for it is of crucial importance to him whether the blow was deliberate or involuntary. If the blow struck was light but deliberate, it has a significance for the person struck quite different from an accidental much heavier blow (H.L.A. Hart, 1968, pp. 182-183).[8]

What matters for responsibility is not merely the effects of one's action but, at least as much, her motivations in acting as she did. Responsibility takes into account what the agent believed herself to be doing and why she did what she did. The idea that one's responsibility is sensitive in some way to one's subjective states is not so much a type of theory but rather a feature of the phenomena for which we must account.[9]

Aristotle, for example, devotes a large portion of Book III of *Nicomachean Ethics* to a discussion of acting under ignorance. Some forms of ignorance excuse, but others do not: "One might give a man a draught to save him, but really kill him" (Aristotle, 1998, p. 52). In such a case the man is ignorant of the effects of his action. Insofar as it was reasonable for the man to think that the draught would save rather than kill he will escape blame despite the harmful effects of his action. But not all ignorance excuses. "We punish a man for his very ignorance, if he is thought responsible for the ignorance… through carelessness; we assume that it is in their power not to be ignorant, since they have the power of taking care" (Aristotle, 1998, p. 60). One can be blameworthy

[8] As quoted by Wallace (1994, p. 124).

[9] I will argue for the stronger claim that an agent's responsibility is completely determined by her subjective states.

for doing something out of ignorance if one ought to have not been ignorant. The psychological states of the agent matter to her responsibility:

> Again, the case of the arts and that of the virtues are not similar; for the products of the arts have their goodness in themselves, so that it is enough that they should have a certain character, but the acts that are in accordance with the virtues have themselves a certain character it does not follow that they are done justly or temperately. The agent also must be in a certain condition when he does them; in the first place he must have knowledge, secondly he must do them from a firm and unchangeable character (Aristotle, 1998, p. 34).

Here Aristotle is restricting the qualities of will that can confer virtue. He is making the plausible claim that for one to be praiseworthy or virtuous one must do the right thing for the right reasons.[10] One must be motivated by the good. For Aristotle quality of will (henceforth QOW) matters to responsibility.

This is in contrast to utilitarian theories of responsibility. These theories hold, roughly, that an agent is blameworthy or praiseworthy to the extent that blaming or praising the agent yields utility. Though some, such as J.C.C. Smart, have held this view it is worth noting that Mill explicitly rejected it and endorsed something closer to a QOW approach: "The motive has nothing to do with the morality of the action though much with the worth of the agent" (Mill, 1979, p. 18).

Some have held that QOW is the only thing that matters at all. For example, consider this famous passage from Kant:

> It is impossible to think of anything at all in the world, or indeed even beyond it, that could be considered good without limitation except a good will…A good will is not good because of what it effects or accomplishes…Even if, by a special disfavor of fortune or by the niggardly provision of a stepmotherly

[10] Though this claim is plausible it is, I believe, false. This is because one's action may fail to be right through no fault of one's own.

> nature, this will should wholly lack the capacity to carry out its purpose—if with its greatest efforts it should yet achieve nothing and only the good will were left-then, like a jewel, it would still shine by itself, as something that has its full worth in itself (Kant, 1998, p. 8).

Kant is claiming that the QOW with which an agent acts is the only thing that is relevant to the question of responsibility or moral worth. It does not matter if the action turns out to be bad due to some external force as long as the agent had the right QOW in acting.

In the twentieth century two especially influential thinkers, Peter Strawson and Harry Frankfurt, have made the notion of quality of will central to their understanding of responsibility.[11] Strawson focuses on what he calls the reactive attitudes, for example resentment and indignation, which are reactions to the quality of will of others:

> 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…The reactive attitudes I have so far discussed are essentially reactions to the quality of others' wills toward us, as manifested in their behaviour: to their good or ill will or indifference or lack of concern (Strawson, 1962, p. 70).

For Strawson, QOW matters to responsibility because our reactive attitudes are constitutive of our practices of responsibility and closely linked with the conditions of responsibility and these attitudes are directly responsive to QOW. In other words, *QOW matters to responsibility because responsibility is about QOW.*

[11] Strawson (1962) and Frankfurt (e.g. 1969; 1971). See McKenna (2005) for a discussion of the insight they share.

Frankfurt, at the end of his essay attacking the principle of alternate possibilities hints at the importance of QOW: "A person is not morally responsible for what he has done if he did it *only because* he could not have done otherwise" (Frankfurt, 1969 p. 10, italics mine). The bare fact that one could not have acted otherwise does not get one off the hook. One is off the hook if *the sole reason* that one acted in some untoward way was that one could not have done otherwise. What matters is the QOW with which one acted.[12] In "Freedom of the Will and the Concept of a Person" he gives an account of QOW, and notes the ways in which different qualities of will might affect one's responsibility.

Frankfurt distinguishes between first and second-order desires. A first-order desire is a desire for some object or state of affairs, for example my desire for a cup of coffee. At any given time we may have many different and conflicting first-order desires. But one of these desires will win out in the sense that it actually moves one to act. The first-order desire that moves one to action is what Frankfurt calls the will. One may also have second-order desires. A second-order desire is a desire for some first-order desire. For example, I may desire the desire to work hard. Of our second-order desires we may distinguish from those that are merely desires for a first-order desire those that are desires that our will be a certain way. He calls these our second-order volitions. For him, an agent acts

[12] This is, I believe, the lesson we should take away from Frankfurt's famous paper. His paper points to what matters in responsibility. And this lesson remains whether his counterexample really does rule out all alternate possibilities.

freely when his will is determined by his second-order volition.[13] This is the quality of will that responsibility requires on a Frankfurtian account.[14]

For example a drug addict may do something bad on the basis of his addiction, say, snatches a purse. The reason that he does this may be because his first-order desire for the drug is irresistible; it is inevitable that that desire would move him to action. But this desire may be at odds with the rest of his commitments. He may have a second-order volition that is contrary to his will. But it is also possible that though his desire for the drug is irresistible, he is entirely content with this state of affairs and would not have it any other way. He may have a second-order volition that he takes the drug. It is intuitive that the willing addict is more blameworthy than the unwilling addict in virtue of the difference in their QOW.[15]

---

[13] Frankfurt is often misread as saying that the mere alignment of one's will with one's second-order volition is sufficient for freedom, but he explicitly denies this: "And it is in the discrepancy between his will and his second-order volitions, or in his awareness that their coincidence is not his own doing but only a happy chance, that a person who does not have this freedom feels its lack" (20-21).

[14] It's worth noting that Frankfurt, in "Freedom of the Will and the Concept of a Person," never actually endorses this view.

[15] On a Frankfurtian account, the presence of a second-order volition to X would enhance responsibility for one's Xing (though Frankfurt in fact never makes this claim). This seems right to me, though I do not believe that the presence of a second-order volition is a necessary condition for moral responsibility. There is also a tension here with what Frankfurt says earlier. Insofar as the willing addict acts on a desire that is irresistible it would seem that his will could not have been what it is *because* of his second-order volition. This is in conflict with his claim that an agent act freely, not when there is mere harmony between the will and a second-order volition, but when the second-order volition causes the will to be what it is. But perhaps Frankfurt is imagining the willing addict's will to be over determined.

In "The Importance of What We Care About," Frankfurt explores the notion of care or concern. When one cares about something that thing *matters* for one or, we might say, one *gives a damn* about that thing. Frankfurt links this notion of caring or giving a damn to that of importance and shows that what we care about plays a significant role in our lives, and in part, determines who we are.

The account of QOW that I seek to develop is indebted to both Strawson and Frankfurt. In my view QOW should be understood by reference to what we care about. When we do things, we do things for reasons, and what reasons we respond to is determined by what we care about. And what reasons others respond to, what motivational structure they have, what their QOW is, matters to us. We care about it. In this way, the reactive attitudes are cares about the QOW of others; they are cares about cares. That is, our practice of holding others (and ourselves) responsible consists in giving a damn about what others (and ourselves) care about.

2. Willing

The notion of the will is one of the most common and most elusive notions in philosophy. For some, it is a prime-mover within us; some special causal power. For others it is merely the result of deterministic forces. There are conceptions of the will that may be difficult to reconcile with a naturalistic world view.[16] The account of QOW that I want to develop does not depend on any controversial metaphysical claims (though it is compatible with those claims). Rather, the will,

[16] I have in mind here agent-causal versions of libertarianism.

or what may be more accurately described as a willing, is simply an action understood internally.

People do things. We vote, we drive over the speed limit, and we hurt each other's feelings. These are actions. Actions should be distinguished from events that merely happen to us. Suppose you push me. The pushing is something that you do but it is something that merely happens to me. It is not an action of mine though it is of yours. Many actions require input from the world. If my filling in a circle with ink is to count as voting the world must be a certain way. There must be an established convention governing the practice and I must be of a certain age and must be at a certain place at a certain time and so on. If a bit of my behavior is to be truly described as my speeding then there must be a law establishing the speed limit on a particular stretch of road, and I must be in a car traveling above a particular speed. If I am to hurt someone's feelings then her feelings must actually be hurt. These are all conditions that are external to me, the agent. But there is a component of action that is internal to the agent. This is the component of action that does not require any input from the world in the sense I have tried to characterize.

Suppose that I hurt your feelings by insulting you. My action can be described as *my hurting of your feelings*. Now imagine another case in which everything is the same except that your feelings are not hurt; perhaps you have a thicker skin than I have supposed. Though my action can no longer be truly described as *my hurting of your feelings* it is importantly similar to my act in the first case. This is because the two actions are qualitatively identical from the first-

person perspective. At the time of action everything was the same from my point of view. I had the same beliefs about what I was doing and I had the same motivations. We can continue to consider cases in which the action is the same from this perspective but the world is increasingly different. For example we can imagine a case in which I attempt to hurt your feelings but fail because you are not before me, there is merely a hologram of your likeness. And we can imagine a case in which though I attempt to hurt your feelings I do not do so because I am a brain in a vat. The component of action that is held fixed across this range of cases is my will.[17] In each case I willed that your feelings should be hurt even though I succeeded in only the first case. Though the world is not held fixed in these cases my willing is. When I speak of an agent's quality of will I am speaking of the quality of her action understood internally in this way.[18]

[17] Some may be hesitant to say that a brain in a vat can act at all. Yet surely everyone will admit that a brain in a vat can will things to occur. I choose to identify actions with willings because I favor a Davidsonian account of action. Actions, such as *his driving of the car*, are simply rediscriptions of willings (See Chapter 3). The adoption of a Davidsonian account of action allows my account to depart less from ordinary language than it would otherwise. For if one were to deny that a brain in a vat acts at all, then my account would imply that agents are only responsible for willings and not for actions. But since on a Davidsonian account actions are willings under descriptions, we can still hold that agents are responsible for actions.

[18] As will become clear later, an action understood internally in this way is not equivalent to the self-conscious perception of action. This is because one's motivations are not necessarily transparent to one.

3. Qualities

When people act they do so for reasons. For example, I may have pushed you because I think you are a jerk or because you were about to step on a rattlesnake. Our motivations in acting matter. I am blameworthy for my action if I pushed you because I think you are a jerk but I may be praiseworthy for pushing you if I was saving you from a snake bite. When I speak of an agent's quality of will I am speaking of the motivational structure that issued in the willing.

QOW is concerned with the reasons that an agent acted as she did; her motives in acting. When some agent acts we can give reasons that attempt to explain the action, reasons that show what the motivations of the agent were. These are one's explanatory reasons or motivating reasons. It is important to distinguish these from one's normative reasons or practical reasons. Suppose that I am at a party and decide to have another drink to ease my anxiety about an upcoming presentation. This drink, however, puts me over the edge and I end up making a fool of myself. We may say that my motivating reason for having another drink was to ease my anxiety. This was why I had the drink. But this is not to say that there was good reason for me to have another. Though there was a motivating reason for my behavior there may have been no normative reason for me to act in that way.

The QOW that some agent actually has is determined by her motivating reasons and these motivating reasons are psychological states of the agent. This is not to deny that normative reasons are relevant. Normative reasons concern, among other things, what QOW one should have. And assessing responsibility

can be thought of as measuring up one's actual QOW with the QOW that one ought to have.

Motivating reasons can come into play at a number of different levels. For example there can be different motivating reasons as to why some agent has some other motivating reason. Suppose that Jack insults a colleague. His motivating reason might be that he dislikes the colleague. But, of course, there might be different explanations as to why he dislikes the colleague. He might, for instance, dislike those who wear sandals. Or he might dislike the colleague because he has witnessed his sexist behavior. These differences are differences in the motivating reasons for the action and hence are differences in the QOW with which he acted. Given the complexity of motivation, QOW and hence responsibility turn out to be similarly complex.

I am thinking of an agent's motivations here in a broad way. This is because I mean to include not only what the agent intends to do but also what she takes the situation to be. I may be motivated to help myself to a second slice of cake at a party. But what I take the situation to be will affect the quality of my willing. My QOW will be different if, for instance, I reasonably believe that there is plenty of cake to go around than it would be if I believed that not everyone has yet had a piece.

People can be motivated in subtle ways. And we can be and often are mistaken about our own motivations. An agent may have a desire to do something while believing that she has no such desire. I may, for instance, believe that I have pursued philosophy due to my selfless pursuit of truth. But it may be that I am

actually motivated primarily by narcissism and my association of philosophy with prestige. As such, agents can be mistaken about the QOW that they have. Given the difficulty, in many cases, of accurately judging agents' motivations, QOW is similarly difficult to judge accurately. As David Shoemaker has rightly pointed out, "human beings are messy and difficult" (2003, p. 117) and our theory of responsibility should reflect this feature. This is also, I think, what Aristotle had in mind when he said that ethics does not admit of the exactness of mathematics.

4. Caring

One's QOW in a given situation is typically determined by what one cares most about regarding that situation. Caring about some thing involves being disposed in certain ways towards that thing. In particular, caring involves cognitive, affective, and volitional dispositions. When one cares about some thing, one is disposed to notice features of the situation that are relevant to the fortunes of that thing. One is cognitively sensitive to factors that can diminish or promote that which one cares about. Similarly, to care about something is to be disposed to certain affective states with respect to the object of one's care. If what one cares about is diminished or harmed, one will typically feel some form of negative affect, and when the object of one's care is promoted in some way one will typically feel positive affect. And finally, we are not passive bystanders with respect to what we care about. If one is in a situation in which one's actions can affect the prospects

of what one cares about, one will be disposed to act so as to promote the enhancement and oppose the diminishment of what one cares about.[19]

If I care about a friend then I will have these dispositions. If I care about my friend then I will typically be sensitive to features of the situation that are relevant to the well-being of my friend. I will be on guard, so to speak, towards things that could harm or help her. If my friend has a tendency to become uncomfortable in social situations, then I will typically be especially sensitive to features of the situation that may cause anxiety in her. I will be looking for clues on her face that she is not comfortable, and be on the look out for uncomfortable situations, say, the boorish acquaintance walking into the party. These cognitive dispositions will be closely intertwined with the affective and volitional ones as well. My noticing features of the situation that have relevance to my friend, will be accompanied by affective responses to the perceived relevant features. My noticing the boorish acquaintance walking up to my friend and the uncomfortable look on her face, will induce in me an uncomfortable feeling. And I will be disposed to act so as to reduce this in both of us. I may, for instance, intercede and ask if she wants to leave the party. Or I may not do anything because I think that it is important for my friend to learn how to deal with situations of this sort because I care about her.

Furthermore, when one cares about some thing this thing is necessarily important to one. This is because one is directly vulnerable to the prospects of the object of one's care. If I care a lot about a sports team, then I am devastated when

[19] Frankfurt (1982), Arpaly (2003), and Shoemaker (2003), have made this point.

they lose and elated when they win. In other words, what happens to the object of my care matters to me because I myself become vulnerable to the fortunes of what I care about. It is in this sense that we identify with the objects of our cares.

One's QOW is typically determined by what one cares most about in a given situation. "If we consider that a person's will is that by which he moves himself, then what he cares about is far more germane to the character of his will than the decisions or choices he makes. The latter may pertain to what he *intends* to be his will, but not necessarily to what his will truly *is*" (Frankfurt, 1982, p. 84). There is an intimate connection between what one does and what one cares about. The connection here is tighter than that between what we decide or intend and what we do. I may intend to act in a particular way but fail because, when push came to shove, it turned out that I cared more about something else.

Shoemaker claims that "what we typically, upon reflection, are motivated to do, in any given situation, depends ultimately on what we care most about, with respect to that situation" (2003, p. 90). Suppose that I am deciding whether to take a job in another part of the country. There are considerations in favor of both taking the job and in staying where I am. I consider the salary, the location, the effect it will have on my loved ones and so on. In trying to decide what to do I am trying to discover which features of the situation I care about most. Do I care more about professional advancement than I do about living in a desirable location? Do I care more about my partner's career than I do my own? This process of reflection will, when successful, reveal what I care most about. And notice that when this revelation is made, so too is my choice. In this way my

decision is an importantly passive process. But since it is based on what I care most about it is a reflection of my true self. Thus, the process is both necessitating and liberating.

Frankfurt makes this point with respect to volitional necessity. Sometimes, one cares so strongly about some thing that one simply cannot act so as to betray that thing. Martin Luther declaring, "Here I stand, I can do no other," was expressing that he cared so much about remaining true to his deepest convictions that he simply could not will himself to recant. Shoemaker expands the scope of this claim to situations in which we reflect on what is important to us. But he believes

> that it is simply false that all my motivated actions depend on things I care about. We all do a variety of things each day that seem to bear no dependence relation at all to our cares; for example, we get out of bed, we scratch itches, we reach for the milk, we change the TV channels, and so on. These are all intentional, motivated actions, explained (rendered intelligible) by reference to certain of our desires, one might say, without any necessary reference to things we regard as important, things whose changing fortunes would tug on our emotional tethers. These are instances in which we act as wantons, then, as unreflective humans who simply do not care what our wills are to be. In such cases we are moved by various impulses, with no real reflection on whether these are the impulses we want to move us, or on whether these impulses flow from what we care about, and the reason is usually that the situation just doesn't warrant that kind of reflection—the situation just doesn't matter (Shoemaker, 2003, p.97).

For Shoemaker what we do in a situation we regard as important is determined by what we care most about. But it is not true that all action is determined by what we care most about. This is because what we do in situations we do not regard as important is not determined by what we care most about. I am somewhat skeptical of these claims. First, note the emphasis on reflection.

Shoemaker seems to be associating caring about something with some kind of reflection or self-awareness with respect to one's cares. This rings false to my ears. Indeed there seem to be many cases in which we simply discover, on the basis of our actions, what it is that we care about. If this is true then whether we reflect on the will we want to have, whether we are Frankfurtian persons or wantons, is a separate issue from whether we care about something. Someone who had no second-order volitions could still care about things it is just that this being (I need to avoid saying person!) is unreflective about her cares.

Not only this, but there is a sense in which the more one cares about something the less refection is necessary or even appropriate. If I care a lot about something, say a loved one, then when I am faced with a choice in which my loved one could be harmed or helped I need not reflect on what I ought to do. Rather, I am simply moved to promote the object of my care. If my loved one is about to be harmed in some way and I am in a position to intervene, then I do. I don't need to stop to reflect on what I care about. This would be, as Bernard Williams has put it, "one thought too many" (1981, p. 18).

Secondly, the way we act in situations we regard as unimportant can be plausibly understood to be determined by what we care most about. What we need to notice is that the object of our care can be something quite particular (e.g. an individual) or quite general (e.g. being happy). The fact that an act of mine is not determined by a sharply aimed care does not mean that there is no general care at work. I am sitting on the couch deciding what to watch on TV. I don't regard the situation as especially important, there is no particular program that I want to

watch, I am just channel surfing and I end up on the food network. This is the kind of case in which Shoemaker would be inclined to say that my watching the food network is not determined by what I care about. Now it may be that there is a specific care that has determined this. I may, for instance, have become more and more interested in cooking as of late but I am unreflective about this. This would explain why I find myself, yet again, watching a cooking show. This is a case in which I discover that I care about cooking. But this is probably not what Shoemaker has in mind. Rather I take it that he is imagining cases in which the fact that I land on the food network is more random. There is no care that determined that I land on this channel rather than another. But actions like these can be understood to be determined by what we care most about. In this case, however, it is not that I care more about cooking than I do about, say, watching a history program. My concern may not push me in one direction or the other. But I do care about being able to watch TV when I have no other pressing commitments. I do care about being able to relax on a Sunday afternoon. I do care about having the opportunity to sometimes sit on the couch and unreflectively channel surf. The frustration of these opportunities would tug on my emotional tether. These examples of unreflective action are simply cases in which I don't particularly care much about how my more general care is satisfied. What we do in situations we regard as unimportant is usually determined by what we care about in the same way that what we do in situations that matter is determined by our cares. One's QOW is typically determined by what one cares most about.

Caring About the Will

Strawson's insight, in "Freedom and Resentment," was that by paying attention to what we are doing when we hold others and ourselves responsible we can get a better grasp on what responsibility is about. "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, p. 62). When we resent another we are reacting to the QOW with which he acted. And when we feel gratitude towards someone who has done something good we are also reacting to the QOW with which she acted. For Strawson, our reactions to the QOW of others and ourselves are constituents of both the practice of and the conditions of responsibility.

What I wish to focus on here is the way in which Frankfurt and Strawson come together through the insight that QOW is what matters for responsibility.[20] In the previous section I argued that one's QOW is typically determined by what one cares about. Strawson's insight was that our own feelings and attitudes, specifically the reactive attitudes of which the practice of responsibility consists, depend upon the QOW expressed towards us by others. My reaction to someone who has deliberately stepped on my toes will be much different from my reaction to someone who has stepped on my toes because he was pushed. In the first, but not the second, case I believe the person to have had a criticizable QOW in acting.

[20] McKenna (2005) argues that the quality of will thesis is the central insight of both Strawson and Frankfurt.

And the fact that my reaction greatly depends on what I take the person's QOW to be shows that QOW matters to me. In other words, I care about it.

The reactive attitudes are themselves expressions of caring; caring about QOW. And since QOW is typically determined by cares[21], the reactive attitudes are second-order cares. To resent another is to care in a particular way about what that other person cared and did not care about. It is to care that the other failed to care more in some way or to care that he cared about the wrong things. For example, if my friend betrays me for financial gain and I resent him for it, then I am caring that my friend cared too much about money and not enough about our friendship. Similarly, when I feel gratitude at the good deed of another, I am caring that that person cared in some admirable or extraordinary way.

Frankfurt saw the importance to responsibility that caring about one's own QOW can make. To be a person is, for Frankfurt, to be the sort of entity that can care about his own will. Strawson saw that we care about the QOW of those with which we interact. Strawson was pointing out that we have, as it were, second-order volitions that are aimed, not at our own wills, but the wills of others. The reactive attitudes are backward-looking reactions to whether that second-order volition has been frustrated, met, or exceeded. The moral obligations we hold others to are forward-looking second-order volitions in this sense.[22]

[21] And can always be redescribed in terms of what we did not care about.

[22] I use the term second-order volition with some reluctance, since for Frankfurt these were a type of desire but I mean to refer to types of cares.

Responsibility is about QOW. Holding responsible involves caring about QOW. It consists in holding someone to an expectation that her QOW be a certain way or fall within a certain range. That is, we expect people to care about morality to some degree. We expect, for example, that in most circumstances persons are to care more about the well being of others than they are the satisfaction of their spontaneous desire to hit another in the face. Persons we think blameworthy are persons we think failed to care enough about acting responsibly. Persons are blameworthy to the extent that they failed to care as they ought to have.

Our practice of responsibility consists in expectations that are directed at QOW. It is important to distinguish this sense of expectation from the epistemic sense of expectation. There is a sense of expectation that means something like "believes it more likely". For example, I may expect some of my students to cheat on the test in the sense that I think that at least one of them will actually cheat. If I were to place a bet I would wager that at least one of my students would cheat. But this is distinct from what we might call the normative sense of expectation. This is the sense in which we hold others to an expectation. Despite the fact that I believe that at least one of my students will cheat, I expect them all to not cheat. That is, I hold my students to this expectation. Were one of them to violate this expectation this would warrant some negative response on my part, such as resentment or indignation.[23]

[23] Wallace (1994) develops this sense of expectation.

We care about the relation between an agent's QOW and our moral expectations. This is what the reactive attitudes are about. When we resent another we are caring that her QOW failed to meet our expectations or demands. The reason that resentment and indignation are attitudes with negative affect is because they involve the frustration of a care. When we feel gratitude we are caring that the person exceeded our expectations. And it is this fact that our care has been promoted that explains why gratitude is an attitude with positive affect. And the large, but often ignored, class of actions which meet our expectations but do not go beyond them are characterized by the typical lack of any reactive attitude. This is because what I care about hasn't been harmed or promoted but remains at the status quo. When I see a driver stop at a red light I am not moved to any reactive emotion. The driver did what I expect of her, no more and no less. Her QOW was ordinary, not extraordinary.

There are many paths to the idea that QOW is what responsibility is about. One way is by reflecting on the phenomena of excuses. Excuses, which serve to diminish blameworthiness, are aimed at QOW. They aim to express that one's QOW was not as criticizable as it may appear. Consider some common excuses: "It was an accident", "I didn't realize that I was doing that", or "It was the result of an epileptic fit".[24] All these excuses attempt to modify the beliefs we have concerning the motivations of the agent. They attempt to show that the agent's

[24] This last plea might not be best thought of as an excuse. Excuses, it might seem, are only excuses for *doing* something and the epileptic might not have acted at all in moving as a result of the fit. However it does seem that the plea "but I didn't do anything at all" is as good an excuse as any.

action did not express the lack of concern normally associated with an action of that type. A taxonomy of excuses would offer a taxonomy of the different qualities of will that one may have.[25] Similar remarks apply to terms of aggravations or, what Michael J. Zimmerman has called, accuses[26] such as: "It was intentional", "he knew full well what he was doing", or "she did it because she is a sadist". These also function, like excuses, by modifying our beliefs about the QOW of the agent, though they do so by making things worse, so to speak. They attempt to express that the agent's QOW was more criticizable than might initially be thought. They attempt to show that the agent cared in the wrong way.

Another path to the idea that QOW is what responsibility is about is from reflection on moral luck. For many, it is strongly intuitive that factors that are external to the agent are irrelevant to responsibility. If the only reason that my attempted murder failed is that a bird flew into the path of the bullet then I am just as blameworthy as I would have been had the attempt been successful. This is because whether the attempt was successful or not was a matter wholly external to me. Whether or not the target is killed I act for bad reasons, I had a criticizable QOW.

For those who reject resultant moral luck, it would be unreasonable to think that one's responsibility can be affected by factors external to one's QOW. One has no choice about factors external to one's QOW and it would be unfair to allow those factors to matter. But, one might think, one also has no choice about

[25] See Wallace (1994).

[26] Zimmerman (1997).

one's QOW. It would be unfair to allow responsibility to be directly sensitive to QOW because one has no choice about that. Some people are born with sunny dispositions while others have the misfortune of being born with stormy ones. It would be unfair to hold one responsible for what one does as a result of one's bad character because one was not responsible for having that character, the objection goes.

One response to this concern is to simply admit that it would be unfair to hold agents responsible for their QOW if they had no choice about it, but to deny that this is so. This is the strategy taken by libertarians. For libertarians, what gives a willing moral content is that it resulted from an indeterministic process (whether this involves agent-causation or ordinary event-causation). This brings up the important point that holding a QOW theory of responsibility does not commit one to compatibilism or incompatibilism, rather it cuts across that debate. Both compatibilists and incompatibilists should be QOW theorists.[27]

For the compatibilist, however, this objection begs the question. QOW theories hold that QOW is *the stuff of responsibility*; it is in what responsibility consists. To object to such a theory by claiming that responsibility is not determined by QOW because one is not typically responsible for having the QOW one has is simply a failure to entertain the hypothesis. QOW theories hold, for example, that to be blameworthy is to have a criticizable QOW. To object then,

[27] Though if the common "tracing" strategy is misguided, as I argue that it is in Chapter 5, then on common libertarian approaches we are responsible for much less than we commonly suppose. This may be taken to support a compatibilist approach.

that one's criticizable QOW is not blameworthy because one is not blameworthy for it is a contradiction on the QOW theory. This is because it amounts to the claim that one's blameworthiness is not blameworthy which is absurd. The objection can make sense only if one assumes that QOW theories are false.[28] The question then becomes whether the claim that one's bad QOW is blameworthy only if one is blameworthy for having that QOW is more intuitive than the QOW theory.

I believe that it is not because it rests on a questionable metaphysics of persons. If it is true, as I believe that it is, that the self consists in some collection of psychological states then to claim that one is responsible for one's psychological states only if one is responsible for choosing to have those psychological states amounts to claiming that one is responsible for what one does only if one is responsible for choosing who to be. As Arpaly notes, "this raises the question of who exactly would be doing the choosing in such a case" (2003, p. 127). This form of self-authorship is, at best, logically puzzling and, at worst, incoherent.[29]

The fact that QOW is what matters for responsibility shows why we should reject resultant moral luck. But it also shows why, for the compatibilists

---

[28] Zimmerman (2008, 175-176) makes this objection. This objection should not be confused with a related one which holds that one's QOW is not criticizable precisely because it is not in one's control or because determinism is true. This objection rests on a controversial understanding of 'criticizable'. The sense of 'criticizable' that I am concerned with is not controversial. It is merely the sense in which some people are citicizable because they care, for example, too much about themselves and too little about others.

[29] See Arpaly (2006, pp. 126-127).

among us, we should accept constitutive moral luck (and for similar reasons, circumstantial luck). This is because luck can play a role in the QOW that one has. This will seem unfair to convinced incompatibilists, but it should be unobjectionable to compatibilists since our place in the causal web is not up to us.[30]

Why do we care as much about QOW as we do? The answer, I think, has to do with our social nature and I imagine that there is some evolutionary story to be told here. Caring as we do about QOW allows us to, among other things, socially navigate the world. It is a mechanism by which we decide who to be around and who to avoid. Furthermore, we care about QOW and the cares of others because we care about *who* people are. This brings up a further point about cares. It is plausible to suppose that the psychological facts about what one cares about are one of the building blocks of personal identity.[31] When we are trying to explain what kind of person someone is we will make reference to her deepest values and commitments. There may be other constituents of personal identity, but it is this one that we seem to care most about.

To summarize: Questions of responsibility are to be settled exclusively by appeal to the QOW with which an agent acted. QOW concerns the quality of an

[30] Though QOW theorists are, for the most part, compatibilists, this is not essential. Indeed, the argument against resultant luck should push both compatibilists and incompatibilists toward a QOW theory. The disagreement then becomes one about the conditions under which a will can have moral content. The incompatibilist claims that the truth of determinism rules out the possibility that a will can have moral content while the compatibilist rejects this.

[31] Or at least what matters in personal identity.

agent's action understood internally. This quality is typically determined by what one does and does not care about. What one cares about is necessarily important to one and it is a fact about humanity that the cares of others matter to us as evidenced by the reactive attitudes. These attitudes are, essentially, second-order volitions. An agent's responsibility for an action is determined by comparing her actual QOW in acting to the QOW she ought to have. I have not provided an account of the QOW that acting responsibly requires. Rather I have tried to move closer to what responsibility is about.

CHAPTER 3

WHAT WE ARE RESPONSIBLE FOR

It is a commonplace that agents can be morally responsible for such things as, for example, *the death of the victim* and *for giving away the surprise*.[32] The primary task of a theory of responsibility, it is thought, is to specify the appropriate relationship one must stand to such things in order to be responsible for them. This is commonly taken to include a control condition and an epistemic condition. Thus, if one is to be responsible for the occurrence of some harm then one must have been in control of the harm and must have known that the harm would result from one's action. I argue that this approach is problematic due to the fact that it attempts to explain the way in which an agent can be responsible for something that is external in a particular sense. Since everything that matters for responsibility is internal, the conditions of responsibility that emerge from this approach are either false or they fail to capture anything of importance.

The problem with this common approach to responsibility is that it attempts to explain how an agent can be responsible for something *external to her mind* or what I will call *an externality*. By this, I mean something that does not make reference to the agent's actual mental states in acting. For example, the consequence *that the dog is run over* is external in this sense since it does not

[32] To forestall any possible confusion, I am thinking of moral responsibility to be a continuum upon which blameworthiness and praiseworthiness are poles. Thus, being blameworthy for something entails being responsible for that thing, but responsibility for something does not necessarily entail being blameworthy for that thing. This sense of responsibility should be distinguished from the mere causal sense with which I will not be concerned.

refer to the agent's motivations in acting. Its occurrence does not necessitate anything about the state of mind of the agent who caused it. Consequences, as typically understood, are external in this sense. Whether actions are external depends on how we understand action.

We can distinguish between act-types and act-tokens.[33] An act-type may or may not be external in this sense. The act-type *stepping on a spider* is external. This is because the occurrence of this act-type does not entail anything about the particular mental states of the agent who did the stepping. While it may, insofar as it entails action, entail intention it does not entail the content of the intention. But the act-type *intentionally stepping on a spider* is not external because its occurrence does entail something about the mental states of the agent. Rather, it is internal to the agent's mind or what I will call *an internality*. Act-tokens, on a common view, are internal in this sense. This is because the act-token *his stepping on a spider* picks out a single event with a unique location in space and time. And the occurrence of that event does entail something about the mental states of the agent, it entails the mental states that the agent actually had in acting. It entails the internal willing that is a constituent of that event. Act-tokens on this view, while not themselves external, can be under external descriptions. *His crossing of the street* is an act-token under an external description. This is because the description

[33] See, for example, Goldman (1970).

itself does not make reference to the mental states of the agent in acting. But what is picked out by that description does have an essential mental component.[34]

Any proposed condition of responsibility should meet the following two constraints: (a) The condition should be such that when we compare a case in which it holds to another in which it does not while holding all other factors fixed there should be a difference in responsibility, and (b) this difference should matter.[35] If the condition fails to meet (a) then it does not capture anything *about* responsibility and thus it cannot be a condition *of* responsibility. If the condition fails to meet (b) then though it may capture something about responsibility it fails to capture anything of importance. All accounts of the conditions under which an agent is responsible for some externality fail either (a) or (b). This is significant since it is commonly thought that the very nature of the project of developing a theory of responsibility consists in developing conditions under which an agent may be responsible for externalities.

The issue is straightforward. Any externality is such that its occurrence or nonoccurrence does not necessitate anything about the state of mind of the agent. And since this is so, for any externality we can imagine cases in which we alter whether it occurs while holding fixed the internal states of the agent. And in such

---

[34] The view I am appealing to here is that of such unifying act theorists as Anscombe (1963) and Davidson (1980) and as opposed to such multipliers as Goldman (1970) and Thomson (1971). But nothing of substance depends on this. The claim I wish to defend only depends on the distinction between internalities and externalities and it will apply whether we adopt a unifier or a multiplier approach to act-individuation. In what follows I will use unifier language but I will note how to rephrase the claims on a multiplier view.

[35] In other words, the difference should itself make a difference.

cases there will be no change in responsibility or at least no change in responsibility that matters. This is because everything that matters for responsibility is internal to the agent. Thus, any account of responsibility for externalities will either be false or it will fail to capture anything of importance. It will be false if we think that what one is responsible for matters. But if we think it is true then we must admit that what we are responsible for does not matter, and thus the proposed conditions will not capture anything of importance. I will further argue that what we are responsible for is important and thus we can only be responsible for internalities.

The Control Condition

It is widely held that agents can be responsible for an externality if they exercise the appropriate form of control over it. Consider John Martin Fisher and Mark Ravizza's account of control. For them, responsibility requires that one exercised guidance control over one's act, omission, or consequence. Guidance control requires the operation of a mechanism that both is the agent's own and is moderately reasons-responsive. The details of the notions of ownership and of moderate reasons-responsiveness are not important for the present purposes. What is important is the fact that the mechanism itself, when the agent's own and moderately reasons-responsive, is internal to the agent. This is because the mechanism is "the process that leads to the action, or the 'way the action comes about'" (Fischer and Ravizza, 1998, p. 38). In cases of responsible agency the

mechanism is “the normal faculty of practical reasoning” (Fischer and Ravizza, 1998, p. 38), and this will include beliefs, desires and other mental phenomena.

For Fischer and Ravizza one is responsible for some externality if it resulted from the agent’s own, moderately reasons-responsive mechanism. Given that the relation between the mechanism, which is internal, and any externality is contingent there are pairs of cases in which we hold fixed the operation of the mechanism but alter whether the externality occurs. This implies, on Fischer and Ravizza’s account, that the agent is responsible for the externality in one case but not in the other. This is problematic because everything that matters for responsibility is internal to the agent’s mind. And this implies either that the account is false or that it fails to capture anything of importance for responsibility.

Suppose that Lee is an assassin and that he aims his rifle at the victim. He pulls the trigger and the victim is shot and killed. There were no responsibility undermining factors; the mechanism that led to the action was the agent’s own and was moderately reasons-responsive. Lee is clearly blameworthy and on Fischer and Ravizza’s account he is responsible for such externalities as *killing* and *the death of the man*. He had control, he knew what he was doing and the victim died. Compare Lee to Harvey who is also an assassin. Harvey aims his rifle at the victim and pulls the trigger. There were no responsibility undermining factors; the mechanism that led to the action was the agent’s own and was moderately reasons-responsive. But after the trigger is pulled a bird flies into the path of the bullet. Harvey does not have guidance control over *killing* or *the death of the man* because there was no killing and no man died. Insofar as everything in

the two cases is the same at and before the pulling of the trigger, Lee and Harvey are blameworthy *in exactly the same way*.[36] I take this as strongly intuitive. Given that everything is the same from the first-person perspective of the agents there is no difference that matters for responsibility. They are both blameworthy to the exact same degree and they are both blameworthy for the exact same reasons.[37] Fischer and Ravizza's account implies that Lee is responsible for killing and for the death of the man while Harvey is not responsible for the killing or for the death of the man. This implies that either the account is false or that it fails to capture anything that matters. That is, if we allow that Lee is responsible for killing while Harvey is not then we must admit that what one is responsible for does not matter for responsibility. But if we think that what one is responsible for does matter then we must admit that the account is false. [38]

The point generalizes to any account of the control one must have in order to be responsible for an externality. Any account of this sort will be either false or

---

[36] One will likely recognize this as a classic case of moral luck (see Nagel [1982] and Williams [1981]). Indeed, the points made here are simply implications of the denial of resultant moral luck.

[37] Note that this does not necessarily imply that they should be punished or otherwise treated in the same way. It just implies that if there is a difference in the punishment that should be given or a difference in how we or they ought to act in light of the action, that difference is not explained by a difference in responsibility. See Zimmerman (2002, p. 562).

[38] Construing the account as a view of responsibility for externalities. This is not to say that the notion of guidance control is not relevant to responsibility for it may be that one can be responsible for an internality only if one expressed guidance control over it. The point here is that the notion of control that is relevant to responsibility cannot extend to externalities.

irrelevant to responsibility. This is because one can fail to have control over the externality without this making any difference that matters for responsibility. Everything that matters for responsibility is internal to the agent.

The Epistemic Condition

Similar remarks apply to the epistemic condition on responsibility. It is commonly held that for one to be responsible for some externality one must stand in a particular epistemic relation to it.[39] Consider John Martin Fischer and Neal Tognazzini:

> [If] Kevin's friends are planning a surprise party for him but they neglect to tell Dan that it's a surprise and Dan subsequently talks openly with Kevin about the party, Dan's ignorance plausibly excuses his behavior. Since he didn't know (and, we suppose, could not have been expected to know) that the party was a surprise, he didn't know that talking openly with Kevin about the party would amount to ruining the surprise. His ruining the surprise is excused because of his impoverished epistemic position. So, it looks like some sort of "epistemic condition" will be a necessary component of any plausible theory of moral responsibility, as well (Fischer and Tognazzini, 2009, pp. 531-532).

For Fischer and Tognazzini, Dan is not blameworthy for his ruining the surprise because he did not stand in the right epistemic relation to it. But suppose that Dan knows that Kevin becomes extremely anxious in social situations. Suppose further that the reason that Dan told Kevin about the party (which he did not know was a surprise) was because he wanted to make Kevin anxious because he doesn't like him. In this case Dan's impoverished epistemic position fails to excuse. In this case, Dan is clearly blameworthy for his behavior.

[39] This idea has its origins in Book III of Aristotle's *Nichomachean Ethics*.

One is likely to respond to such a case by claiming that though Dan is not responsible for *his ruining the surprise*, he is responsible for *his making Kevin anxious* because he knew that telling Kevin about the party would have this effect.[40] But this can be shown to be false when one considers cases where one's belief fails to amount to knowledge. Suppose that Dan believes that telling Kevin about the party will make him anxious and he does this for vicious reasons, yet Dan has no justification for this belief. As it happens his unjustified belief is true. In such a case Dan is blameworthy *in exactly the same way* as he is in the case where his belief amounts to knowledge. This has led some, like Gideon Rosen, to the view that what matters for the epistemic condition for responsibility is not knowledge but true belief: "Whenever we start with a case in which the agent knows the wrong-making features of his act and then consider related cases in which the agent truly believes that his act has these features but fails to know that they do (either because his belief is not justified or because it lacks some other knowledge-relevant feature), the agent will be every bit as blameworthy as he was in the original case" (Rosen, 2008, p. 597). For Rosen, responsibility requires, not that one know of the wrong-making features of one's action, but simply that one has true beliefs regarding those properties. But true belief is vulnerable to the same considerations that knowledge is. When we start with a case in which the agent truly believes his act to have particular wrong-making features and then

[40] Note that this strategy precludes one from holding a unifier account of act-individuation, unless one holds that agents are responsible, not for acts, but for act-descriptions. This is because on a unifier account *his ruining the surprise* and *his making Kevin anxious* are simply different descriptions of the very same event.

consider a related case in which the agent falsely believes his act to have those features, the agent will be every bit as blameworthy as he was in the original case. We can imagine a case in which Dan believes that telling Kevin about the party will make him anxious and he does so, but Kevin never becomes anxious because he has recently started taking anti-anxiety medication. In this case Dan is still blameworthy *in exactly the same way* as he is in the other cases. This suggests that what matters for the epistemic condition for responsibility is not knowledge, or true belief, but simply what the agent believes or ought to believe. And note that once we make this shift, we are no longer providing an account of the epistemic relationship between an agent and some externality since what an agent believes or ought to believe are internal to her mind.[41]

Any account of the epistemic relation one must stand to some externality in order to be responsible for it will either be false or will fail to capture anything of importance to responsibility. The epistemic requirement on responsibility that matters does not consist in some epistemic relationship an agent must stand to some externality. It is a condition that is wholly internal to the agent.

What We Are Responsible For

The lesson of the above discussion is that what matters for responsibility is internal to the agent. I also believe that what we are responsible for matters. Thus, I believe that what we are responsible for must be internal.

[41] This is not to say that there is no sense of "ought to believe" that is external. But the sense of "ought to believe" that is tied to responsibility is necessarily internal.

This is in contrast to Michael J. Zimmerman. Zimmerman does believe that what matters for responsibility is internal to the agent because he rejects resultant moral luck. He believes, for example, that the successful and unsuccessful assassins in the above case are blameworthy *to the exact same degree*. But he believes that they differ in the scope of their blameworthiness. One is blameworthy for the death of the victim while the other is not. It is this that leads him to say that "degree of responsibility counts for everything, scope for nothing" (Zimmerman, 2002, p. 568).

I've argued that it is inconsistent to think both that agents can be responsible for externalities and that what one is responsible for matters. Zimmerman, who also recognizes this inconsistency, resolves it by accepting the former and rejecting the latter. I, however, reject the former and accept the latter. I do this because it is more plausible to think that what one is responsible for is relevant to one's responsibility than it is to think that one can be responsible for externalities. The heart of this disagreement concerns the following principle:

> *The only things for which one can be responsible are those things in virtue of which one is responsible.*

I find this principle extremely plausible.[42] To claim that S is blameworthy for X is not, contra Zimmerman, to make an uninformative claim that is irrelevant to anything that matters for responsibility. It is to claim that X explains S's blameworthiness; X is the reason that S is blameworthy. And these reasons are internal.

[42] David Copp (1997) also defends this claim.

If we accept these claims, then we cannot be responsible for externalities. Thus, we cannot be responsible for the consequences of our actions since they are external.[43] This is because consequences are distinct events from the actions that caused them. In order for a given consequence to result from some action the world must cooperate in some way and this is a matter external to the agent. Though we cannot be responsible for the consequences of our acts we can be responsible for our acts understood as act-tokens.[44] This is because act-tokens are individuated finely. They are events with essential internal properties. An act-token is internal though it may be under an external description (e.g. *his killing of the dog*). But it is important to note that the external description is not the responsibility relevant one. The descriptions that are responsibility relevant are the internal descriptions. We can also be responsible for act-types though only those that are internal. The act-types that we can be responsible for are those picked out by the internal descriptions of the act-token that are responsibility relevant. For example, suppose a man shoots a child for malicious reasons. The consequence *that the child is killed* is not something for which the man is

[43] This should not lead one to think that cases of negligence are problematic on this account. One might think that in cases of negligence we hold the agent responsible for the consequence but that there is no associated internality of the kind I find relevant to ground responsibility. But there is a relevant internality that will ground responsibility, specifically, the internality *her acting negligently*. Acting negligently typically involves not taking due care in acting, and taking care is an internal notion.

[44] On a unifier account act-tokens are necessarily internal. But on a multiplier account some act-tokens are internal and some are external. The external act-tokens on a multiplier account are the internal act-tokens under an external description on a unifier account.

responsible. This is because whether that consequence occurs is an external fact. Similarly, the man is not responsible for the act-type *killing a child* since this is also external. He is responsible, however, for the act-token *his killing of a child*. But the description *his killing of a child* is not the responsibility relevant one. Rather it is the description *his trying to kill a child for malicious reasons* that is relevant. Both descriptions pick out the same event but it is the latter that, in an important sense, explains his responsibility; it is this description that describes the features of the act that make it blameworthy.

Conclusion

The traditional approach to theorizing about responsibility leads to a distorted picture. Insofar as the traditional approach attempts to explain how an agent must be connected to an externality in order to be responsible, it implies that if the connection is severed so too is responsibility. But this is clearly false. We've considered cases in which an agent fails to have control over some externality but this failure fails to excuse. And we've considered cases in which an agent fails to be epistemically connected to some externality but this failure also fails to mitigate. The traditional approach to responsibility leads us to prematurely cut off our inquiry into the responsibility of agents. Even if we know that some agent failed to have control over or failed to know about some externality she could, for all we know, be fully responsible. She may be just as responsible as she would have been had she freely and knowingly brought about the externality. What we need to know in order to know her responsibility is not her relation to some

externality. Rather we need to look inward. We need to look at her motivations and her beliefs about what she was doing and the care she took in so doing. We need to know about her quality of will.

CHAPTER 4

BLAMEWORTHINESS AND WRONGNESS

Many have held that agents can be blameworthy only for morally wrong acts. This chapter argues that if one holds this claim, and makes a plausible assumption about rightness and wrongness, one is forced to accept an implausible view about moral responsibility. Instead, this claim should be rejected. Agents can be blameworthy for acts that are not morally wrong. The chapter is divided in the following way: Section 1 puts the point in terms of three initially appealing, but jointly inconsistent propositions. Section 2 motivates the significance of noting this inconsistency by pointing out a number of theorists who have held, or at least flirted with, all three propositions. Section 3 argues that the best way to resolve the inconsistency is to reject the claim that blameworthiness requires wrongdoing. Section 4 considers and rejects a natural alternative to the first proposition. Section 5 suggests a further replacement for the first proposition, one that garners more intuitive support than either of the previous two considered.

1.

The following three propositions are jointly inconsistent:

(1) One is blameworthy for some action only if that action is wrong.

(2) Factors external to one's mental states at the time of the action can affect whether one's action is wrong.

(3) Factors external to one's mental states at the time of the action cannot affect whether one is blameworthy for the action.[45]

If (1) and (2) are true then it is possible that some external factor that affects whether one's act is wrong can affect whether one is blameworthy making (3) false. If (1) and (3) are true then there is no possible external factor that can affect whether one did wrong making (2) false. And if, as it will be argued, (2) and (3) are true then the connection between blameworthiness and wrongness can't be as direct as (1) suggests.

(1) does seem rather intuitive. It is appealing to think that what explains the blameworthiness of an agent is the fact that she did wrong. There seems to be a very close relationship between the notion of blameworthiness and the notion of wrongness and (1) seems to be a plausible candidate for that relationship.

The sense of externality appealed to in (2) and (3) can be thought of in terms of supervenience. A is external to B just in case there can be a change in A without a change in B. That is, A is external to B if and only if A does not supervene on B. With this consideration in mind, (2) says that some factor that does not affect the agent's mental states at the time of the action could affect the moral status of the action. This is plausible, and in fact, most theories of right and wrong have this feature. The clearest example of a theory of right and wrong, what I will call a *normative theory*, that accepts (2) is consequentialism. For example, according to a utilitarian version of consequentialism, overall utility

[45] Note that (1)-(3) could be put in terms of praiseworthiness and rightness as well.

determines whether some action is morally right or morally wrong. Overall utility is a factor that is external to the agent's mental states, in the sense that there could be a change in overall utility while holding fixed the mental states of the agent. One could have the purist intentions but cause a catastrophe, and for that reason one would have done wrong, according to this normative theory.

Action-based deontological normative theories also have a commitment to (2). For example, a deontological theory that holds that it is morally wrong to kill an innocent will be committed to the claim that factors external to the agent's mental states can affect rightness and wrongness. This is because whether, in fact, an innocent is killed is a matter external to the mental states of the agent in acting. We can easily imagine cases in which we hold the mental states of the agent fixed but alter whether a death occurs. Insofar as these theories hold that whether an actual death occurs can affect whether one has done wrong, then they will be committed to (2).

Virtue based normative theories also suggest a commitment to (2). It is plausible to think that insofar as the notions of rightness and wrongness play a role in the virtue ethicists thinking, it is something along the following lines: The right action, in a given circumstance, is the action that would be chosen by the virtuous person in the circumstance. And, perhaps, the wrong action in some circumstance is that action which would be avoided by the virtuous person in the circumstance. If this is in the right ballpark, then it does seem that virtue based normative theories are committed to (2), for the rightness or wrongness of some

action is determined by a kind of idealization test that does not appeal to the actual mental states of the agent in acting.

(3) makes the plausible claim that it is some aspect of an agent's mental life that determines responsibility such as her intentions, desires, or beliefs. It says that factors that are external to one's mental states at the time of action cannot affect whether and to what degree one is blameworthy for the action. For example, if the reason that one's attempted murder is unsuccessful is that a bird flew into the path of the bullet then one is just as blameworthy as one would have been had there been no bird present. (3) denies the possibility of resultant moral luck or luck in how things turn out.[46] It says that the degree to which one is blameworthy for some action depends upon one's mental states at the time of the action. The truth of (3), however, is consistent with the possibility of other forms of moral luck. (3) says that one's mental states are what matter for responsibility. If luck plays a role in determining the quality of one's mental states, say, those of a sinner or a saint, then that luck can play a role in the degree to which one is responsible.[47]

(3) is the insight shared by so-called quality of will theories of responsibility. These theories all hold that what matters for responsibility are particular psychological states of the agent. These theories do often differ about which mental states they hold to be central. Strawson (1962), for example, held

[46] See Nagel (1982).

[47] For a discussion of quality of will and constitutive moral luck see Arpaly (2003, Chapter 5).

that it was the attitudes expressive of good or ill will toward ourselves or others. For some what matters is that one's second order volition aligns with one's will (Frankfurt, 1971), for others it is a choice (Wallace, 1994), or one's evaluative commitments (A. Smith, 2005), or one's concern (Arpaly, 2003; 2006). But though these theories may disagree about which mental states are central to responsibility, they all hold that responsibility is determined by one's mental states at the time of the action. Putting this idea in slogan form, we might say, "responsibility just is in the head."

2.

Despite the appeal of each of the three propositions (1)-(3), they cannot all be true. But despite this inconsistency a number of theorists have held that they are all true, or that it is possible that they are all true. Many theories of moral responsibility do not purport to answer the question of which acts are right and which acts are wrong. That is, they attempt to remain neutral concerning which normative theory is correct. Presumably, these theorists want their accounts to be consistent with most normative theories such as those mentioned above. But some of these theorists clearly hold (1) and (3) and so would be forced to deny (2) and thus their accounts of responsibility would only be compatible with a minority of normative theories.

For example, H. Smith proposes the following account of blameworthiness:

S is blameworthy for performing act A if, and only if:
1. Act A is objectively wrong,
2. S had a reprehensible configuration of desires and aversions,

And

3. This configuration gave rise to the performance of A (Smith, 1983, p. 556).

The first condition, it should be obvious, entails (1). It says that a necessary condition of blameworthiness is that the action must be wrong. The second and third conditions express a mental states based approach to blameworthiness: "To blame someone is to criticize that person for some, perhaps short-lived, psychological state insofar as it manifests itself in action" (Smith, 1983, p. 556). Presumably to blame another involves believing that the other is blameworthy and so blameworthiness, on Smith's account, would seem to involve criticizable psychological states.

While Smith does not explicitly endorse a claim such as (3), she does indicate sympathies toward such a view. She distinguishes between a broad and narrow view of the factors that can affect blameworthiness. "On the narrow view only psychic factors contribute [to an agent's blameworthiness]; on the broad view consequences of one's actions contribute as well" (Smith, 1983, p. 568). And later she says, "Personally, I suspect no framework will be found to support the broad view" (p. 570). I take it that Smith would accept (3) on these grounds. Since she accepts (1) and (3) her account of blameworthiness is incompatible with any normative theory that entails (2).

Another example can be found in the work of Arpaly: "I take it as an intuition that in order for me to be blameworthy for an action, it has to be the case that the action is wrong" (2006, p. 91). So she accepts (1), that a necessary condition for blameworthiness is that one has done wrong. But she is also a self-

described quality of will theorist: "A person is praiseworthy for taking a morally right course of action out of good will and blameworthy for taking a morally wrong course of action out of lack of good will or out of ill will" (Arpaly, 2006, p.15). Whether an agent acts from good or ill will is determined by whether the agent was motivated by moral concern. Arpaly "take[s] concern to be a form of desire" (2003, p. 84). And since an agent's desires are clearly internal to her mental states, so too must be her moral concern. Given this it is plausible to think that she would accept (3).[48] For her what matters for responsibility is the quality of will with which an agent acts and an agent's quality of will is internal to her mental states.

She does not give or defend an account of the right-making and wrong-making features of action and seems to think that her account does not stand or fall with any particular normative theory. She says:

> Which reasons exactly are moral reasons is not a question I can deal with here, as the moral reasons to perform an action are the same reasons that make the action right, and what exactly makes an action right is a question that Kantians, utilitarians, Aristotelians, and others are still debating (Arpaly, 2003, p. 72).

[48] I believe she would also be moved to accept (3) based on considerations of resultant moral luck. Given her lengthy discussion attempting to assuage worries that a quality of will account such as hers will be committed to constitutive moral luck (2003, Chapter 5), I take it she would readily deny resultant moral luck (which is, I believe, more intuitively unpalatable than constitutive luck). And denying resultant luck seems to bring a commitment to (3).

Arpaly's account of praiseworthiness and blameworthiness includes (1) and (3). And so she cannot accept any normative theory that entails the truth of (2), such as those mentioned in the passage.[49]

Perhaps the most striking example can be found in Kant, who is often taken as the paradigm example of a theorist who thinks that morality is immune to luck. Though the interpretive issues here can be difficult, my point is that on a natural and plausible reading Kant can be seen to hold (1)-(3). Kant allows for the possibility that one does the right thing for the wrong reasons. Indeed, in stressing the importance of acting from duty, he focuses on cases in which agents acts dutifully but not from the motive of duty. For example, he considers shopkeepers who price fairly from different motives. One prices fairly because it is what duty calls for, and for this reason his action has moral worth. Another prices fairly but does so because it will increase profits. Since this second shopkeeper does the dutiful thing, but not from a motive of duty, his action does not have moral worth.

[49] Another way to see the tension in Arpaly's account is by considering examples in which we hold the mental states of the agent fixed but alter the moral status of the action. Suppose that in one scenario an agent successfully murders his victim. The action is morally wrong. But suppose that in a second scenario we hold the mental states of the agent fixed but alter the moral status of the action. Suppose, for example, that the gun misfires and no one is killed and the action is not wrong according to the utility calculus. Given this Arpaly must say that the second agent is not blameworthy since his action was not wrong. But I take it she would not want to say this since it is plausible to think that because the two agents had the same mental states at the time of the action that they acted with the same quality of will and hence that they must be equal with respect to responsibility. The only options she has are to either deny that the actions have a different moral status (that is, to deny (2) and take a stand on which normative theory is correct), or to hold that the agents did act with different qualities of will (that is to deny (3), and to hold that quality of will is external to one's mental states). Neither option seems desirable.

This suggests that the normative theory that Kant has in mind is committed to (2). The shopkeeper examples show that the criteria by which we judge whether an act was dutiful is external to the agent's mental states. Since the shopkeeper examples are examples in which the fact that the action was dutiful is held fixed but the mental states of the agents are altered (whether the agent acted from a motive of duty or prudence), then whether an action is dutiful or not is not determined by the mental states of the agent in acting. So Kant appears to be committed to (2).[50]

And as Herman notes, he also holds (1): "And when we say that an action has moral worth, we mean to indicate (at the very least) that the agent acted dutifully from an interest in the rightness of his action: an interest that therefore makes its being a right action the nonaccidental effect of the agent's concern" (Herman, 1993, p. 6). Here Herman is (correctly) attributing to Kant (and endorsing) (1).[51] She is saying that a necessary condition for an action to have moral worth (or perhaps for an action to be praiseworthy or for an agent to be praiseworthy in light of her action) is that it must be in accordance with duty (i.e. it must be morally right). So Kant accepts (1). Herman's quote also provides more evidence that Kant is committed to (2), for the requirement that the rightness of

[50] Assuming that we can think of the dutiful action as the right action. This the reading favored by Arpaly: "Recall Kant's Prudent Grocer, who prices his merchandise fairly because a reputation for honesty tends to increase his profits. Despite the Prudent Grocer's unimpressive motive, Kant never denies that the grocer does the right thing or that he performs the action required of him by duty. In this sense, Kantians clearly care about results" (70).

[51] Strictly speaking it is (1)'s corollary which says that only right acts can be praiseworthy.

one's action is the "nonaccidental effect of the agent's concern" seems to imply the possibility that the rightness of one's action could be accidental. This is the case with the prudent shopkeeper. The fact that he gets it right is an accident since he is motivated by a concern for profit. And if this is true, then the criteria by which we judge rightness must be external to one's mental states. So Kant accepts (1) and (2). And, of course, he famously accepts (3):

> A good will is not good because of what it effects or accomplishes…Even if, by a special disfavor of fortune or by the niggardly provision of a stepmotherly nature, this will should wholly lack the capacity to carry out its purpose—if with its greatest efforts it should yet achieve nothing and only the good will were left-then, like a jewel, it would still shine by itself, as something that has its full worth in itself (Kant, 1998, p. 8).

3.

Inconsistency could, of course, be resolved by giving up any one of the three propositions. Williams (1981) and Nagel (1982), for example, would be inclined to give up (3). Slote (1996) would give up (2).[52] The best solution, however, is to give up (1).

In support of (3) consider two scenarios: In the first scenario an assassin carefully aims his gun at his target and pulls the trigger. The victim is killed. In the second scenario the assassin carefully aims his gun and pulls the trigger. But a bird flies into the path of the bullet before it reaches the intended victim. We can imagine that the two scenarios are indistinguishable from the assassins' point of

[52] On Slote's agent-based view, whether an act is right or wrong is determined by the mental states of the agent: "An agent-based approach to virtue ethics treats the moral or ethical status of acts as entirely derivative from independent and fundamental areatic (as opposed to deontic) ethical characterizations of motives, character traits, or individuals" (Slote 1996, 83).

view up until the bullet has left the barrel. Given this, it seems that their blameworthiness is the same in virtue of the similarity of their mental states at the time the trigger is pulled.[53] The presence or absence of the bird is a matter of luck external to mental states of the agents, and so it seems irrelevant to blameworthiness.

Suppose that the target in both cases was a Mother Teresa type figure who helps those in need and does a great deal of good for the world. In the first case, where the assassination is successful, the world suffers a great loss and utility is certainly not maximized. For this reason, the action would be wrong on utilitarian grounds. But in the second case, there was no assassination and so the action cannot be wrong for the same reasons. Furthermore, we might suppose that the bird that flew into the path of the bullet was actually carrying a new and extremely contagious and dangerous strain of bird flu. Had the bird avoided the bullet, suppose, it would have infected other birds which in turn would have infected humans and caused a global pandemic. Given this, it would seem that we might judge the action as right on utilitarian grounds. Utility was maximized by shooting the bird. It seems, then, that we have some reason to accept (2). Accepting (2) does not require that we hold that consequences are everything, but only to admit that they can make a difference. To deny (2) is to disallow the

[53] This is consistent with the claim that the assassins are blameworthy for different things. One is blameworthy for his killing, while the other is blameworthy for his unsuccessful attempt. (3) makes a claim about *whether and to what degree* one is blameworthy, not *that for which* one is blameworthy. Though in Chapter 3 I do argue for the claim that *that for which* one is blameworthy is internal to one's mental states.

possibility that actual consequences external to the agent's mental states can matter at all to the question of right and wrong.

Furthermore, to deny (2) and to hold that the factors that determine rightness and wrongness are facts about one's mental states threatens to dissolve the distinction between rightness and wrongness, and praiseworthiness and blameworthiness. If the moral status of one's action is determined by facts about one's mental states, such as the reasons that one did what one did, then it would not seem possible to do the right thing for the wrong reasons. This is because the factors that determine whether one did the right thing would be facts about the reasons for which one acted. But it also seems that whether one is blameworthy or praiseworthy is determined by facts about the reasons for which one acted. Denying (2), then, might make it impossible to do the right thing for the wrong reasons (or to do the wrong thing for the right reasons) and for this reason, it might conflate rightness and wrongness with praiseworthiness and blameworthiness. [54]

(2) and (3), then, are more plausible than (1) and so (1) should be rejected. Others have rejected (1), though for different reasons.[55]

---

[54] One might hold a view in which rightness and wrongness are determined by some set of mental states of the agent at the time of the action, and that blameworthiness and praiseworthiness are determined by some other distinct set of mental states of the agent at the time of the action. Such a view might escape the proposed objection but I find a view like this to be implausible.

[55] See Haji (1998), Zimmerman (1997), and Scanlon (2008). For a related discussion see Parfit (forthcoming). Parfit seems to be one of the few theorists who has recognized that holding (1) might bring a commitment to moral luck.

4.

The first proposition, that blameworthiness requires wrongdoing, seems patently false when one considers cases in which an agent tries to do wrong but accidently gets it right. It seems, in such cases, that the agent is blameworthy despite a failure to do wrong. Consider a case in which a nefarious doctor wants to kill his patient and injects the patient with what he has every reason to believe is a deadly poison in the belief that he is committing a moral wrong. But, as it turns out, the substance is not a poison but a medication that in fact saves the patient's life. Surely the doctor is blameworthy, but on many normative theories his action is not wrong.[56] Examples like this cast doubt on (1). They also show the need to include an epistemic element into the account of the relationship between blameworthiness and wrongness. Some have suggested that what is required for blameworthiness is not that one actually does wrong, but that one *believes* that one is doing wrong. Consider:

> (1a) One is blameworthy for some action only if one believes the action to be wrong.[57]

This is consistent with (2) and (3) since, plausibly, the beliefs one has about what one is doing necessarily affect the mental states that one has. It also

---

[56] One might object that the doctor's unsuccessful attempt was morally wrong. It can plausibly be replied that this is to confuse wrongness with blameworthiness.

[57] This view has been advocated by Haji (1998, Chapter 8). Zimmerman and Parfit advocate a similar view which holds, roughly, that freely acting in the belief that one is doing wrong is sufficient for blameworthiness. This seems doubtful in virtue of Huck Finn type cases in which one can't help but do the right thing despite one's grossly mistaken beliefs about the demands of morality.

seems to account for our intuitive judgment concerning the above example. One of the reasons that the doctor is blameworthy for injecting his patient with a life saving medication is that he thought that he was actually killing the patient. He acted in the belief that he was doing wrong and this seems to explain, in part, the judgment that he is blameworthy. Yet (1a) fails for it lets too many agents off the hook. History is rife with examples in which people do the wrong thing in the belief that they are doing right. It is plausible to suppose that a good number of those people are in fact blameworthy. Suppose that historians announce that they have conclusive proof that Hitler believed he was doing the right thing in waging genocide. Should such a discovery call into question the blameworthiness of Hitler? This seems implausible.[58] The problem with linking blameworthiness directly to belief, as (1a) does, is that sometimes peoples' beliefs about what is right and wrong are entirely unreasonable. Ignorance is not categorically exculpating.

5.

Is there another candidate that does better than (1) or (1a)? One of the problems with (1) was that it lacked an epistemic dimension. It did not take into account the beliefs of the agent. But though (1a) did take into account epistemic considerations it failed too for it lacked a normative dimension. An account of the

[58] For a discussion of why some mistaken beliefs speak ill of an agent see Arpaly (2003, Chapter 3).

relationship between blameworthiness and wrongness needs to take into account both epistemic and normative considerations. Consider the following:

> (1b) One is blameworthy for some action only if one ought to believe the action to be wrong.

This is also consistent with (2) and (3) since, as is argued below, what one ought to believe, in the relevant sense, is a fact about how the world seems from the agent's point of view and this is surely a fact about the agent's mental states. Unlike either (1) or (1a), (1b) seems to get the right results in all of the cases discussed. In each of the cases in which the agent is blameworthy it is also the case that the agent ought to have believed the action to be wrong. One reason that the doctor is blameworthy for injecting his patient with the medication is that he ought to have thought that he was doing wrong. Indeed, in this case, he actually did believe what he ought to have believed. But even if he lacked the belief that he was doing wrong he would still be blameworthy as long as it is the case that he ought to have thought he was doing wrong and there were no other responsibility inhibiting factors. And Hitler is blameworthy for similar reasons. Even if he lacked the belief that he was doing wrong, he ought not to have. He ought to have believed that mass extermination is morally wrong. This is, in part, why he is blameworthy. And finally, reconsider the example involving the two assassins. Though the presence or absence of a bird turned out to affect the moral status of

the actions, it could not have affected what the agents ought to have believed. In both cases the assassins ought to have believed that they were doing wrong.[59]

It is important to say a bit more about the sense of ought appealed to in (1b). It has been argued that if blameworthiness requires wrongdoing, and whether one does wrong can depend on factors external to one's mental states, then whether one is blameworthy can depend on factors external to one's mental states. But this is implausible so we should reject the claim that blameworthiness requires wrongdoing. The same problem would arise if the sense of ought appealed to in (1b) could depend on factors external to the agent's mental states. But this must be denied. The factors that determine what an agent ought to, in this sense, believe supervene on the agent's mental states. That is, there could be no change in what an agent ought to, in this sense, believe without a change in the agent's mental states. This is because what one ought to believe depends on one's actual beliefs and one's evidence and these are facts about one's mental states.[60]

One might object that (1b) cannot be correct in light of the following kind of example. Suppose some doctor is faced with the choice of administering a substance to a patient. The doctor has decisive evidence for the belief, (φ), *that the substance is medicine*. But the doctor comes to have the irrational belief, (ψ),

---

[59] Assuming, of course, that there was no evidence of the presence of the bird at the time the trigger was pulled.

[60] This is not to say that there is no sense of "ought to believe" or evidence that can depend on factors external to one's mental states. But this sense, so long as it is external, cannot be the sense that is tied to responsibility. That is, the ought that is tied to responsibility, must be determined by factors internal to the agent's mental states (this is just to say that (3) is true).

that the substance is poison. Acting on (ψ), the doctor tries to kill his patient by giving her the substance. This might be a case in which the doctor ought to believe that his action is right (since (φ) is supported by the evidence) but yet he is blameworthy, and hence, it might be thought to show (1b) to be false.

The fact that (φ) is supported by the evidence does not make it false that the doctor ought to think his action is wrong. This is because there are two different considerations that can affect what one ought to think regarding the normative status of one's action. The first consideration concerns one's actual beliefs about what one is doing. The second consideration concerns the beliefs about what one is doing that one ought to have. Failing to meet at least one of these standards is a necessary condition for blameworthiness. "Ought to believe wrong", then, should be understood in terms of a disjunction. Consider:

> (1b*) One is blameworthy only if one ought to believe one's action is wrong in virtue of either (a) one's actual beliefs about one's action or (b) the beliefs about one's action that one ought to have.

(1b*) is an explication of (1b) rather than an alternative to it. (a) concerns the content of one's actual beliefs about one's action. If one believes that one is now stepping on a baby, then one ought to think that one is doing wrong. This is true in virtue of the content of one's actual belief. (b) appeals to the content of the beliefs one ought to have. Suppose Jim is babysitting a 5 year old, Alexa. Alexa's mother has told Jim that Alexa is deathly allergic to peanuts and has given Jim a note explaining this along with emergency contacts. But Jim's attention is elsewhere. He is thinking about the football game that is about to start. Alexa's

mother leaves and Jim turns on the game. A couple of hours later Alexa wants lunch. Jim, preoccupied with the game, throws together a peanut butter and jelly sandwich and gives it to Alexa. Alexa dies of an allergic reaction. Even though what Jim actually thought he was doing was not wrong, he ought to have thought that what he was doing was wrong. This is true in virtue of the content of the belief he ought to have had: "I am giving Alexa a PB and J which puts her in mortal danger since she is deathly allergic to peanuts." The fact that Jim gave Alexa a PB and J despite the fact that he ought to have thought it to be wrong is a fact about his quality of his will (which is a fact about his mental states), and this, in part, makes him blameworthy.

What is it for it to be the case that one ought to believe that one's action is wrong in virtue of some belief? What features of the belief make it true that one ought to think that one is doing wrong? "Ought to believe" can be understood in terms of "it is reasonable to expect one to believe." (1b*), then, can be rewritten using the notion of reasonable expectations in place of 'oughts':

> (1b**) One is blameworthy only if one ought to believe that one's action is wrong, either because (a) it is reasonable to expect one to infer from one's actual beliefs about one's action that one's action is wrong or because (b) there are some beliefs about one's action it is reasonable to expect one to have and those beliefs are such that it is reasonable to expect one to infer from them that one's action is wrong.

If one ought to believe one's action is wrong in virtue of one's actual belief about the action, then it is reasonable to expect one to infer from the assumption that the

belief is true to the conclusion that the act is wrong.[61] For example, it is reasonable to expect a competent adult to infer from "I am now stepping on a baby" to "I am doing wrong". And thus if one believes that one is now stepping on a baby one ought to think that one is doing wrong. If one ought to think that one's action is wrong in virtue of the beliefs about the action it is reasonable to expect one to have, then there are some beliefs it is reasonable to expect one to have and those beliefs are such that it is reasonable to expect one to infer from the assumption that they are true to the conclusion that the act is wrong. Reconsider the example of Jim and Alexa. There is some belief about Jim's action it is reasonable to expect him to have: "I am giving Alexa a PB and J which puts her in mortal danger since she is deathly allergic to peanuts." And this belief is such that it is reasonable to expect Jim to infer from its truth to the conclusion that what he is doing is wrong. In this way, Jim ought to believe that what he is doing is wrong.

The notion of expectation, as it is used here, is not a purely epistemic notion. It is not the sense of expectation that is based on likelihood. I might expect one to act in some way that I believe is very unlikely. For example, I may expect my students to do their reading, even though I believe that they won't. And should some student fail to do the reading that I expect her to do, this may warrant some reaction on my part, for the student, we might say, did not hold up her end of the bargain. The sense of expectation used here is conceptually bound to the notion of

[61] For the sake of simplicity I am ignoring complications involving probability. To allow for these kinds of cases we would need to replace 'wrong' with 'not expectably best'. See Parfit (forthcoming) and Sepielli (2009).

response. Holding another to an expectation involves, at least, believing in some way or other that a breach of the expectation warrants some response or actually being disposed to respond in some way to such a breach. The claim, then, that it is reasonable to expect one to believe something, amounts to claiming that were one to fail to have that belief certain responses would be warranted. The expectations appealed to in (1b**) are of a moral kind and so the reactions that they imply are moral reactions. When one fails to meet a reasonable moral expectation this may warrant some reaction such as indignation or resentment, or it may impair one's relationship with other moral agents.

It has been claimed that one is blameworthy for some action only if one ought to believe that the action is wrong. This amounts to the claim that one is blameworthy for some action only if it is reasonable to expect one to think one's action is wrong. The conditions under which one is blameworthy for some action entail particular conditions concerning the normative standards applicable to one's epistemic situation.

It should now be clear how to respond to the putative counterexample to (1b) in which the doctor has evidence to which he is unresponsive that his action is right. The doctor has an actual belief such that it is reasonable to expect him to infer from its truth to the conclusion that the action is wrong. That is, it is reasonable to expect him to infer from his actual belief "I am giving this patient poison," to the conclusion "I am doing wrong". This is what makes it true that he ought to think his action is wrong. The fact that he has evidence to which he fails to respond for the belief that the substance is medicine does not undermine the

above consideration. The fact that he has decisive evidence for the belief that the substance is medicine does not make it false that he ought to think his action is wrong.

Furthermore, note that (1b), understood in this way, is compatible with the truth of (2), and (3). This is because the requirement for blameworthiness put forth by (1b) is a requirement that is internal to the agent's mental states in the sense that the factors that determine what it is reasonable to expect one to believe about one's action necessarily affect one's mental states. Jim's mental states, in giving Alexa a PB and J, would have to have been different had it been the case that it was *unreasonable* to expect him to believe that she was allergic to peanuts. Suppose Alexa's mother never mentioned Alexa's allergy to him. If this were the case then his will in giving her a PB and J would not have displayed a negligent lack of concern. What is, in fact, reasonable and unreasonable to expect one to believe about one's action necessarily affects one's mental states in acting. (1b), then, is consistent with (3), the claim that blameworthiness is an internal notion, and (2), the claim that wrongness is an external notion.

This chapter argued that propositions (1)-(3) are jointly inconsistent, and that we would do best to give up (1), the claim that blameworthiness requires wrongdoing. This should be replaced with (1b), the claim that blameworthiness requires that one ought to believe that one is doing wrong. This should not seem like a radical suggestion. This relation, between blameworthiness and wrongness, is captured by the common response we direct at those who plea ignorance: "You should have known better."

CHAPTER 5

RESPONSIBILITY, TRACING, AND CONSEQUENCES

Quality of will (henceforth QOW) accounts of moral responsibility hold that an agent's responsibility for an action is completely determined by some aspect of the agent's mental life at the time of the action. Such accounts are often objected to in light of "tracing cases." These are cases in which an agent acts in such a way at an earlier time that he is unable to express the appropriate moral agency at some later time, and yet he is, intuitively, morally responsible at that later time. If such examples succeed they show QOW accounts to be false. My project here is twofold. First I will argue that the QOW theorist can respond in a plausible way to the tracing cases. The strategy involves holding that the agent is responsible for his earlier action, but not the event to which the action led. Secondly, I will argue that not only is this a viable option for the QOW theorist, but that it is the only option for theorists of moral responsibility in general. This is because the tracing approach is a strategy employed in order to account for an agent's responsibility for the consequences of her actions but, I will argue, agents cannot be responsible for the consequences of their actions.

On QOW accounts of responsibility an agent's responsibility for some action is determined by the agent's QOW in acting. Different theorists have put forth various accounts of the relevant qualities of will. For Peter Strawson (1962), it is the attitudes expressive of good and ill will toward ourselves or others. For Harry Frankfurt (1971), it is the way in which an agent identifies with his willing

by way of a second-order volition.[62] For others, it is a choice, or one's concern, or one's evaluative commitments.[63] Though there is disagreement about which components of an agent's mental life determine his responsibility they do all hold that responsibility is determined by some aspect of the agent's mental life at the time of action. QOW accounts can be characterized by the slogan "responsibility just is in the head."

For example, consider Frankfurt's (1971) influential account. Frankfurt distinguishes between an agent's first and second-order desires. A first-order desire is simply a desire for some object or state of affairs. At any given time we may have a host of different and often conflicting first-order desires. But one of these desires will be effective in the sense that it will move one to act. For Frankfurt, the first-order desire that moves one to action is one's will. Secondly, we often have desires that take as their object another desire, what he calls an agent's second-order desire. I may, for example, desire the desire to work long hours. Of an agent's second-order desires we can distinguish those that are not merely the desire to have some desire but the desire that some particular first-order desire be effective. These he calls second-order volitions; an agent's desire that his will be a certain way. On this view an agent is responsible when there is an alignment between the agent's will and his second-order volition.

---

[62] McKenna (2005) argues that the central insight from both Frankfurt (1969) and Strawson (1962) concerns the Quality of Will Thesis: "Being morally responsible and legitimately holding morally responsible are to be settled exclusively in terms of the moral quality of the will with which an agent acts" (172).

[63] See Wallace (1994), Arpaly (2003), and Smith (2005) respectively.

According to QOW accounts, such as Frankfurt's, having the appropriate QOW is necessary and sufficient for responsibility. There are, then, two ways in which one may object to such an account. On the one hand, one may attempt to come up with cases in which an agent expresses the appropriate QOW and yet, intuitively, is not responsible thereby showing that QOW is not a sufficient condition for responsibility. This is what the common "manipulation cases" attempt to establish. Frankfurt's account, for example, is often objected to since it seems that an agent could be manipulated, through hypnosis, brainwashing, or direct stimulation of the brain, into having the appropriate alignment of his will and his second-order volition. For example, perhaps a team of psychologists have implanted in some agent both an effective desire to murder and a second-order volition that his desire to murder be his will. But, it is claimed, surely the agent is not responsible for his subsequent murder given the manipulation.[64] I will not attempt to respond to such objections here.[65]

Rather I will focus on objections of the second kind. Those that attempt to show that QOW is not a necessary condition of responsibility. These are cases in which it is intuitive that the agent is responsible despite lacking the appropriate QOW. The drunk-driving case is the stock example used in this context. Consider an agent who freely and knowingly becomes inebriated at a party and then proceeds to drive home and subsequently runs over a child. The agent, imagine,

[64] See, for example, Fischer and Ravizza (1998), Haji (1998), Mele (2009), and Pereboom (2001).

[65] Though see McKenna (2008) for a response.

was so drunk at the time of the accident that he was not capable of forming second-order volitions, yet he is, intuitively, responsible for killing the child. Such cases, if they are successful, show that QOW is not a necessary condition of responsibility. Instead it is thought that the notion of tracing, which can be traced back to Aristotle,[66] is an essential component of a theory of responsibility and that we must *trace back* from the untoward event to some previous choice in order to ground the agent's responsibility for the event. For example, John Martin Fischer and Neal Tognazzini "do not see how a theory of moral responsibility could adequately handle the range of drunk-driving cases, Martin Luther cases, and manipulation cases without some sort of tracing component; tracing just seems both highly plausible and theoretically indispensable" (2009, p. 553). Manuel Vargas has argued that the notion of tracing is significantly more problematic than has generally been assumed. He does note that structuralist accounts (such as Frankfurt's, a type of QOW account) need not rely on tracing, though he thinks this counts against them: "The absence of any role for tracing in structuralist accounts seems at least as problematic as the troubles caused by tracing in non-structuralist accounts. Pedestrian cases (so to speak) involving drunk drivers, as well as a range of somewhat more esoteric manipulation cases, seem deeply problematic for structural theories" (Vargas, 2005, p. 287). In what follows I will suggest how a QOW account of moral responsibility can adequately handle the drunk driving cases without appeal to tracing.

[66] See *Nicomachean Ethics*, Book III [1113b20-1114a27].

Responsibility, as I am using it here, is taken to be the extent to which an agent is blameworthy or praiseworthy for her action. Thus being blameworthy entails being responsible but being responsible does not entail being blameworthy (for one may be praiseworthy). I will focus on cases of blameworthiness in the remainder of the chapter, yet one could substitute cases of praiseworthiness and the point would remain. My argument will apply to responsibility in general so long as it is the case that a difference in blameworthiness for some action entails a difference in responsibility for that action and that no difference in blameworthiness for some action entails no difference in responsibility for that action. Because I believe that blameworthiness simply is a form of responsibility I find these principles extremely plausible, yet I do admit that there may be other ways of thinking about these issues. The sense of moral responsibility appealed to here should not be confused with either the legal or the merely causal senses with which I will not be concerned.

Before continuing I want to say a bit more about the context of the dialectic in which this debate occurs. The broader issue at stake is whether responsibility is an essentially historical notion.[67] A historical notion is one in which historical factors play a metaphysical determining role. For example, consider the notion of a counterfeit work of art. We can imagine two versions of Munch's "The Scream." Suppose that one is the original while one is a counterfeit. But also imagine that the counterfeit is perfect in that the two are indistinguishable on the basis of their physical properties. Despite this it is still the

[67] See, for example, Fischer and Ravizza (1998, Chapter 7).

case that one of the two is the original and one is the counterfeit. The features of the paintings that make it true that one is the original and the other is a counterfeit are features about their respective histories. One of the two has a history that involved being painted in the late 19th century, while the other does not. But there is no physical property of either that indicates this difference. The difference relates solely to their histories. The notion of a counterfeit work of art is a *genuinely historical notion*.

Critics of QOW theories generally hold that responsibility is genuinely historical. This is because they think that one must invoke historical conditions in order to adequately account for the manipulation cases and the tracing cases. A key feature of QOW theories, then, is that they are ahistorical or "current time-slice" accounts. The debate about whether the tracing cases are problematic for QOW theories occurs within a broader debate about whether responsibility is historical. If, as I hope to show, tracing cases are not problematic for QOW accounts, then this undermines part of the motivation to think that responsibility is a genuinely historical notion.

The chapter is divided in the following way: First, an overview of the strategy is presented. This strategy holds that in the tracing cases the agent is really blameworthy only for that which occurs at the point to which the tracers trace. In the drunk driving case, this is commonly taken to be the choice to become drunk. Some worries are briefly raised about the use of the standard drunk driving case as the paradigm in which to invoke tracing and a modified version is offered. It is then argued that we should not appeal to tracing in order to

explain blameworthiness for consequences because, despite appearances, agents cannot be blameworthy for the consequences of what they have done. Having argued that consequences are not related to blameworthiness in the way that many think, I address the way in which they are related. I argue that the consequences of one's actions are only *epistemically relevant* to blameworthiness. The chapter closes by considering and responding to some possible objections.

1.

Consider the standard drunk driving case: Jack freely and knowingly becomes inebriated at a party and chooses to drive home. On his way home he runs over a child in the street. Surely Jack is blameworthy for the death of the child. But he lacked *responsibility grounding agency* at the time of the accident, for he was too drunk. Responsibility grounding agency (RGA), as it is used here, is a placeholder for whatever capacities or other features it is that distinguishes moral agents from non-moral agents. This is often taken to be some form of control and may involve reasons responsiveness, libertarian free will, alignment of one's will with one's second-order volition, or some other condition. (RGA therefore includes, but is not limited to, the features to which the QOW theorist appeals).

The reason Jack is blameworthy for the death of the child, the proponent of tracing holds, is that we can *trace back* from the unfortunate event to his decision to get drunk. And at this point he both possessed the relevant agency and satisfied an epistemic condition; he could reasonably be expected to believe that his action would put others at mortal risk.

In the above drunk driving case the tracer holds that Jack is blameworthy for the death of the child since we can trace back from that consequence to a time at which he expressed RGA and satisfied the relevant epistemic condition. The natural place to which to trace is, according to the tracer, to the point at which he decided to get drunk.[68] Tracing is thought to be especially necessary when we compare Jack to Jill, who also ran over a child but who did so only because her Shirley Temple was secretly spiked. Intuitively, Jill is not blameworthy despite the fact that factors at the time of the consequence are relevantly similar to those in Jack's case. It is thought that tracing is necessary to explain these asymmetrical judgments of blameworthiness. Only in Jack's case can we trace back to a time at which he possessed RGA and the consequence was expectable.[69] The strategy to be explored here involves holding that Jack is really blameworthy only for his becoming drunk.[70] Since blameworthiness is not ascribed to a point at which RGA is lacking we need not invoke tracing.

The drunk driving cases and others relevantly like them have the following features:

(a) The agent did not exercise RGA during the consequence C.

---

[68] See, for example, Fischer and Tognazzini (2009, p. 532).

[69] Jill fails to meet the epistemic condition since we can stipulate that she was reasonably unaware of her drunkenness and hence the consequence was not expectable.

[70] As will become clear, this assessment of the standard drunk driving case is mainly for illustrative purposes. I have concerns about the use of this standard example which are addressed in the next section. The more guarded claim is this: Insofar as it is unproblematic to apply tracing to the standard drunk driving case, the agent is really blameworthy only for that to which the tracers trace.

(b) The agent did exercise RGA during the action A.

And the proponent of tracing holds that since

(c) A is related to C in the right way and the agent could reasonably be expected to believe that A would result in C,[71]

(d) The agent is blameworthy to degree *d* for the consequence C.

The strategy advocated here, however, involves denying (d). This may seem counterintuitive but it seems less so when it is emphasized that while (d) is false the following is true:

(d*) The agent is blameworthy to degree *d* for the action A.[72]

Notice, this is a crucial point, that the above strategy involves changing only *that for which* the agent is blameworthy. We need not change *whether and to what degree* the agent is blameworthy. In Jack's case the pre-theoretical intuition is that he is blameworthy for the death of the child to some degree *d*. We can still hold that he is blameworthy to that degree *d*, we just change that for which he is blameworthy. It seems that the feature of the example that is intuitively powerful concerns the judgment *that the agent is blameworthy to this degree*. On the strategy suggested here we can retain that judgment.

[71] Or the agent could reasonably be expected to believe that A would make C more likely. In what follows I bracket off issues concerning probability.

[72] For some it may not be obvious that blameworthiness comes in degrees. But consider that excuses are rarely fully exculpating, but more often mitigating. This shows that blameworthiness is a scalar notion.

2.

In the standard drunk driving case it can be quite unclear that the agent lacked RGA at the time of the consequence. For, after all, the agent was still able to drive. I suspect that in most actual cases of drunk driving there is not a complete lack of RGA. Consider two scenarios: The first is the standard tracing case in which an agent runs over a child as a result of her freely and knowingly becoming inebriated. Contrast this with a second agent who also has become inebriated freely and knowingly and runs over a child. But suppose that the second agent actually *aims* at the child in the street. It is plausible to think that the degree of control required to aim at a child is guaranteed by the fact that the agent is driving at all. Insofar as the second agent aims at the child it seems that the second agent is more blameworthy than the first. If this is true, that there is a difference in blameworthiness between the first and second agents and we can hold all the pre-driving factors fixed, then the difference must be explained by some factor at the time of the driving. And this factor seems to be a particularly objectionable expression of agency on the part of the second agent. If this is true, then the drunk driving cases are not ones in which the agent completely lacks RGA at the time of the accident, and hence, there is no need to invoke tracing.

It can also be unclear in the standard drunk driving case to where we ought to trace. We need to trace to a point at which the agent expresses RGA and the consequence is expectable. But, in the standard case, there may be an inverse relationship between these two conditions. The more plausible it is that the agent expresses RGA at some time (e.g. during the choice to have that first drink) the

less plausible it becomes that the consequence is expectable at that time. And the more plausible it becomes that the consequence is expectable at some time (e.g. after the fifth drink) the less plausible it becomes that the agent expresses RGA at that time.

To ensure that these vagaries are not hindering our assessment of tracing consider a modified version of the example. Suppose that rather than freely and knowingly getting drunk, Ernie freely and knowingly takes a pill that gives him a sense of euphoria (without impairing RGA) but will cause him to suddenly fall into a deep sleep a couple of hours later. He takes the pill, knowing its effects, has a good time at the party and later gets into his car and starts to drive home unreasonably thinking that he can make it home before he falls asleep. A bit later he falls asleep, slumped over the steering wheel. After a few minutes of the car fortuitously staying on the road, it veers into a playground and runs over a child. Bert's case is similar, however, the pill is surreptitiously slipped into his Shirley Temple (and suppose that Bert reasonably does not realize that he has been drugged, to him it just seems like he is having a good time). He too gets into his car and starts driving home. He eventually falls deeply asleep due to the pill, slumps over the steering wheel and later the car runs over a child. In these cases, as before, it is intuitively clear that Ernie is blameworthy while Bert is not. But it is also clear, as it may not have previously been, that at the time of the death neither agent expressed *any* of the relevant agency, for they were both fast asleep. And Ernie's driving of the car before he fell asleep was such that he both exercised RGA and the consequence was reasonably expectable. In this variation

of the drunk driving case it is clear that at the time of the consequence the agent lacks RGA and there is also a clear prior time at which the agent does have RGA and, in Ernie's case, satisfies the epistemic condition.

3.

The tracing strategy holds that Ernie is blameworthy for the death of the child in virtue of the fact that we can trace back, from the point at which the death occurred to a point at which he did express the relevant agency and the consequence was expectable. In the modified example this would be Ernie's driving of the car before he fell asleep. We cannot do this in Bert's case, for he does not satisfy the epistemic condition, and so Bert is not blameworthy for the death of the child.

The strategy argued for here holds that in the tracing cases, the agent is really only blameworthy for that which occurred at the point to which the tracers traced. Ernie is blameworthy for his driving of the car but not for the death of the child. This is because the death of the child is a consequence of Ernie's action and agents cannot be blameworthy for the consequences of their actions. Consider the following argument:

(1) If some thing is not that in virtue of which one is blameworthy then one cannot be blameworthy for that thing.

(2) The consequences of one's actions are not that in virtue of which one is blameworthy.

(3) Therefore, one cannot be blameworthy for the consequences of one's actions.

The first premise, which has the ring of analyticity, says that the only things that one can be blameworthy *for* are the things that *make* one blameworthy.[73] A denial of (1) will be open to the charge that it makes blameworthiness *unfair*.

This can be seen by appeal to the notion of resultant moral luck.[74] To many, the idea of resultant moral luck is unpalatable. If the only reason that one's attempted murder was unsuccessful is because a bird flew into the path of the bullet, it seems that one is just as blameworthy as one would have been had the attempt been successful.[75] The idea is that blameworthiness ought not to depend on factors that are, from the agent's perspective, a matter of luck. But notice that there are two ways in which resultant luck might be claimed to affect blameworthiness. One way is that luck might be thought to be capable of affecting *whether and to what degree* one is blameworthy. In this way, luck could affect whether one is blameworthy and could also make one more or less blameworthy. I believe luck of this kind should be rejected. Another way in which resultant luck

[73] David Copp (1997, p. 453) also makes this claim, though he puts it to a different use as part of a defense of the principle of alternate possibilities.

[74] The classics on the topic are Williams (1981) and Nagel (1982). Resultant moral luck occurs when external factors that affect how one's action turns out affect one's responsibility. While Williams and Nagel point out this phenomenon it is Zimmerman (1987) who coins the term "resultant moral luck".

[75] The claim that the successful murderer and unsuccessful murderer are equal with respect to blameworthiness is compatible with the claim that they differ with respect to legal culpability. It very well may be that moral responsibility is not the only relevant factor for determining legal responsibility.

might be thought to affect blameworthiness has to do, not with whether and to what degree one is blameworthy but with *the scope* of one's blameworthiness.[76] That is, resultant luck might be thought to be capable of affecting *that for which* one is blameworthy. I believe both these forms of resultant luck should be rejected.[77] (1) claims that the scope of one's blameworthiness is not subject to resultant luck.

The second premise may be thought to do the heavy lifting, but before considering it in more detail some terminological clarification is called for. Here, and in what follows, 'consequence' will be used somewhat as a term of art. 'Consequence' will be taken to mean an event or state of affairs (causally related in the appropriate way to an action of an agent) under a description that makes no

---

[76] Zimmerman (2002) makes this distinction. He holds that though luck cannot affect whether and to what degree one is blameworthy he is committed to the view that luck can affect the scope of one's blameworthiness. For example, he holds that though the successful and unsuccessful murderers are blameworthy to the same degree, the successful murderer is blameworthy for more things (e.g. the death of the victim).

[77] That is, we should take the implications of the denial of resultant luck "to their logical conclusion" (Zimmerman 2002; p. 559). Zimmerman takes himself to be doing this, but he is not since he still allows for the scope of one's blameworthiness to be subject to luck. In order to make this form of luck seem less objectionable Zimmerman claims that scope counts for nothing. On his view, then, the claim that S is blameworthy for X counts for nothing. But this is false; scope counts. To claim that S is blameworthy for X is not to make an uninformative claim, it is to claim that X makes S blameworthy.

I will not here provide much of an argument for the claim that resultant luck should be rejected. Domsky (2004) argues that we should reject resultant luck, in part, because the intuitive pull to accept it is a result of optimistic and selfish biases. In other words, we tend to judge negligence that unluckily results in harm worse than negligence that luckily does not because we are implicitly optimistic that we will be lucky and we selfishly allow the unlucky to shoulder the moral burden. According to Domsky, "believing in moral luck is much like clucking like a chicken after seeing a hypnotist" (463).

reference to the mental states of the agent in acting. For example suppose that one flicks a light switch, which causes a prowler to be alerted. *One's flicking of the light switch* (which can be redescribed in various ways) is one's action and results in the consequence *that a prowler is alerted*. *That a prowler is alerted* is a consequence, in the relevant sense, since it is an event (appropriately related to the action) under a description that makes no reference to the mental states of the agent in acting. [78] Note that though this is a more precise and perhaps idiosyncratic conception of the notion of a consequence it captures the majority of occurrences of our more common and ordinary notion, and it also applies to the consequences in the examples of the tracing proponents.[79]

Insofar as the description of the consequence includes no reference to the mental states of the agent in acting, for any consequence there is some possible fully exculpating excuse and some possible fully aggravating accuse. An accuse (rhymes with excuse) is Zimmerman's (1997) term for a consideration that shows a bit of putatively non-blameworthy behavior to be blameworthy. Suppose, for

---

[78] See Davidson (1971). One possible redescription of *one's flicking of the light switch* is *one's alerting of the prowler*. And since, on Davidson's view, both are simply different descriptions of very same event, on my view, the agent can be responsible for the action under this description. One, perhaps odd, result of my account is that though the agent cannot be responsible for the consequence *that a prowler is alerted* the agent can be responsible for *his alerting of the prowler*. The reason has to do with the truth conditions of these two events. The truth conditions of the consequence, *that a prowler is alerted*, are independent of the mental states of the agent. But the truth conditions of the action, *the agent's alerting of the prowler*, are not independent of the mental states of the agent since that action is identical to *the agent's intentional flicking of the light switch* which does make essential reference to the agent's mental states. Note that though I do favor a Davidsonian conception of action nothing in my argument requires it.

[79] See, for example, Fischer and Ravizza (1998, Chapter 4).

example, that Jones saves the life of Smith. Typically this is not the sort of action that warrants blame. But suppose we learn that though in fact Smith was saved this was just a lucky accident, and that Jones actually intended to kill Smith. Given any consequence C there is some possible scenario in which C occurs but the agent is not at all blameworthy and some other scenario in which C occurs and the agent is maximally blameworthy. There are also cases in which one's blameworthiness is not at all affected by the presence or absence of the consequence. To see that this is so imagine cases in which the consequences that occur are not, from the agent's perspective, reasonably expectable. For example suppose some agent is related to the world in such a way that each action of the agent triggers some bad (or good) event. The case can be imagined such that the agent is reasonably unaware of this unfortunate (or fortunate) situation and, from her subjective perspective, lives a virtuous (or vicious) life. The idea is that it seems that blameworthiness and praiseworthiness are dependent upon the mental life of the agent and the relation between this mental life and its consequences is a contingent one.

It may be objected that to assert that for any given consequence there is some possible fully exculpating excuse and some possible fully aggravating accuse is simply to beg the question with regard to the second premise of the above argument. Two points: First, consider an example of the sort described in more detail. In Orson Scott Card's science fiction novel *Ender's Game*, the protagonist, Ender, is a student at an elite military Battle School. He quickly moves to the top of his class after showing many tactical talents and is promoted

to attend Command School. For his final exam at Command School he plays a video game which simulates a large battle. Ender's ships are greatly outnumbered by the alien fleet orbiting around their alien planet. Displaying his willingness to win at all costs he decides to use a special weapon which destroys the planet and with it all the alien ships and some of his own. It is later revealed (spoiler alert) that the game was not a mere simulation, but that his actions actually controlled the movement of troops and that the "simulated" events actually took place culminating in alien genocide. This is a case in which the normal correlation between expectable consequences and actual consequences is severed. For Ender the expectable consequences of his actions just concern the game and the few other students that he was playing with (who also believed that they were only playing a game). In such a scenario it is clear that Ender's blameworthiness (if any) is completely determined by his subjective states. So long as it is the case that he reasonably believed that he was merely playing a game he cannot be blameworthy in virtue of the alien genocide. This is compatible with the claim that he is blameworthy for some other thing, say, for displaying a certain kind of brutality in his game playing. But Ender has a fully exculpating excuse with regard to the destruction of the alien species: "I was reasonably unaware that my actions would result in genocide or any other harm". Now consider a twist on Ender's story in which he reasonably believes that the game is real and that his actions will lead to genocide. But also suppose that in fact the game is only a simulation. In such a scenario Ender is just as blameworthy as he would have been had it been the case that the game was real and he knew this to be so. What

Ender's blameworthiness turns on is what he can be reasonably expected to believe he is doing (and the reasons for which he is acting), not what his actions in fact cause. We would not, for instance, let him off the hook were he to plea that luckily no harm was caused. When considering cases like Ender's it is clear that it is the *expectability* of particular consequences rather than their actual *occurrence* that matters for blameworthiness. Since the relation between the consequences that are expectable and those that actually occur is contingent, it will always be possible for any actual consequence to imagine a fully exculpating excuse and a fully aggravating accuse. This shows that the occurrence of some consequence of one's action is not that in virtue of which one is blameworthy.

Secondly, though it is possible some will insist that the occurrence of some consequence does necessarily affect one's blameworthiness, those who wish to deny the existence of resultant moral luck should accept the points made here. This is because whether some consequence results from a given action will always be subject to an element of luck since the relation between the consequences that are expectable and those that actually occur is contingent. The idea, mentioned above, is that what seems to matter for blameworthiness is the agent's perception of what she is doing. This is not to say that good intentions always excuse, for one's perception of what one is doing is subject to normative standards. But if one took due care and yet the world failed to cooperate, this does not reveal anything criticizable about the agent. And to be blameworthy is to be criticizable in a particular way. Insofar as one is uncomfortable with the idea of

resultant moral luck one should also be uncomfortable with blameworthiness for consequences.

The occurrence of some consequence is not that in virtue of which one is blameworthy. Consequences with some particular property are neither required nor sufficient for blameworthiness. For any consequence C and any agent's blameworthiness B we can imagine cases in which we hold C fixed and alter B, and also cases in which we hold B fixed and alter C. This shows that one's blameworthiness is independent of the consequences of what one has done, in the sense that the occurrence of some consequence is never that in virtue of which one is blameworthy. Now if one can only be blameworthy for that in virtue of which one is blameworthy, then one cannot be blameworthy for the consequences of one's actions.

4.

Agents cannot be blameworthy for the consequences of their actions. In what way, then, are actual consequences relevant to blameworthiness? Consequences are relevant to blameworthiness only insofar as they provide evidence concerning the mental life of the agent but it is aspects of this mental life for which agents are blameworthy. The consequences of one's actions are only *epistemically relevant* to blameworthiness. Some thing is only epistemically relevant to blameworthiness when it can provide evidence about an agent's blameworthiness but is not constitutive of it. In the tracing cases, the event that spurs the tracing (a consequence) is only epistemically relevant. In these cases the agent is really

blameworthy only for that which occurred at the point to which the tracers traced. In the modified version of the drunk driving example Ernie is really blameworthy only for *his driving of the car* (an action) before he fell asleep and the child was hit. And since, in Bert's case, he was reasonably ignorant of the fact that he had been slipped the drug, this explains why we don't find him blameworthy. Blameworthiness is not ascribed to a point in which RGA is lacking, and hence, we need not invoke tracing.

Suppose Greg shoots his pistol at the ceiling to celebrate at a wedding. The bullet ricochets off the ceiling and hits and kills the bride. A consequence of Greg's action is *that the bride is killed*. This fact is epistemically relevant to the blameworthiness of Greg. This is because the fact that the bride was killed as a result of his shooting of the pistol provides evidence that he was negligent, a fact about his mental states. The fact that the bride was killed provides evidence that Greg failed to be appropriately moved by the fact that, from his perspective, it was reasonably expectable that his shooting of the gun would risk harm. When it is thought that Greg is blameworthy for killing the bride it is inferred from the occurrence of that consequence that his willing (the pulling of the trigger) had certain qualities.[80] It is thought that his willing displayed a willingness to put

[80] So long as the reader will permit me to do a bit of psychological speculation. The ultimate point though, is that this is what people ought to think whether or not they actually do so**.** Furthermore, this is not mere speculation. There is a strong empirical case to be made for the claim that the tendency to judge negligence that results in harm more harshly than equal negligence that does not result in harm is explained by an implicit judgment that the negligence in the harm case is worse (what has been called hindsight bias). For a discussion of moral luck and epistemic bias see Royzman and Kumar (2004) and Domsky (2004).

others in danger. It is this in virtue of which he is blameworthy. Notice, too, that this willingness is independent of the occurrence of the consequence. In an alternate scenario in which the bullet ricochets in a different direction and hits someone else or even does not hit anyone Greg is just as blameworthy because he had the same mental states at the time of action; his willing had the same qualities.[81]

But suppose Greg pleads the following excuse: "I was given this gun, which I was told was loaded with blanks by the usher and was instructed to fire it at the end of the ceremony to celebrate." Insofar as we believe that he is telling the truth we may think that he is actually not blameworthy at all.[82] Successful excuses modify our beliefs about the mental states of the agent at the time of the action. Sometimes excuses do appeal to the consequences of what one has done. But the success of these excuses turns on what can be inferred about the mental states of the agent from the given consequence. The relationship between the occurrence of some consequence and the blameworthiness of an agent is not one of necessity.

Returning now to the case of Ernie, the consequence *that the child is killed* is only epistemically relevant to the question of his blameworthiness. Often, when

---

[81] One might wonder in what sense the occurrence of the consequence provides evidence of blameworthiness since the view defended here holds that Greg is blameworthy to the same degree and for the same thing whether or not the consequence occurs. The answer is that the occurrence of a bad consequence provides evidence of blameworthiness proportionate to the degree to which consequences of that type are typically expectable.

[82] Assuming we don't think that he ought to have been more certain that the gun only contained blanks before firing away.

such a consequence obtains there is a blameworthy agent behind it, someone who did not appropriately respond to the relevant considerations. But this only occurs in the case of Ernie and not in the case of Bert. It is this that explains why Ernie and not Bert is blameworthy. Since what Ernie is actually blameworthy for was his driving (in which he had RGA) after freely and knowingly taking the drug we need not invoke tracing to explain the asymmetrical judgments of blameworthiness.

The tendency to think that agents are blameworthy for the consequences of their actions is based on a confusion between epistemic and metaphysical considerations. Some consequence-types are such that there is a high correlation between their occurrence and an agent with particular criticizable mental states that ground blameworthiness; fatalities resulting from drunk driving, for example. But this correlation is explained by the fact that the consequence is typically expectable in these kinds of cases. And it is the expectability of the consequence, rather than the occurrence that explains the agent's blameworthiness.

The tracer suggests that what explains the agent's blameworthiness for the consequence is that there is some prior time at which

(a) The agent expressed the relevant agency and

(b) The agent could reasonably be expected to believe that that expression of agency would bring about the consequence.

But notice that once it is clear that (a) and (b) are satisfied, the question of blameworthiness is settled. All of our reasonable moral criticisms of the agent will concern what the agent has done at that prior time. It may be that the

occurrence of some consequence *motivates us* to investigate questions concerning (a) and (b). And it may be that had the consequence not occurred, it would not have *occurred to us* to ask these questions. But, of course, (a) and (b) can occur in cases in which the consequence does not. We do often make certain inferences about the quality of the agent's willing from the occurrence of some consequence but it is mistake to see the relevance of the occurrence of the consequence as metaphysical rather than epistemic.

Let me reiterate that the claim is that the *occurrence* of some consequence is not directly relevant to blameworthiness. This is not to say that no considerations involving consequences are directly relevant. For it is surely the case, as stated above, that the *expectability* of consequences is directly relevant.[83] The claim here is that with respect to blameworthiness the occurrence of a consequence can, at best, provide evidence that the agent did not appropriately respond to the fact that, from her perspective, the consequence was expectable. The consequences of an agent's action are epistemically relevant in that they can provide a window into her mental life at the time of the action.

5.

Here two objections are considered to the claim that agents cannot be blameworthy for the consequences of what they have done. First, one might

[83] One might think of the difference between a theory of the right and a theory of responsibility is the differing roles they give to consequences. A theory of the right is often concerned with the occurrence of consequences, while a theory of responsibility ought to be more concerned with the expectability of consequences.

attempt to save the possibility of blameworthiness for consequences in the following way: One might grant that when the consequence of one's action that actually occurs is not expectable one cannot be blameworthy for the consequence (for, of course, one has a good excuse). But one might hold that when the consequence that occurs is expectable this allows for blameworthiness for the consequence (for one lacks such an excuse). In reply, such a move strikes me as *ad hoc*. The only motivation for this response is that it saves the pre-theoretical intuition that agents can be blameworthy for the consequences of what they have done. But in scenarios in which actual consequences and expectable consequences do line up, it is the fact that the consequence is expectable that makes it the case that the agent is blameworthy, not the occurrence of the consequence. And that for which one can be blameworthy is limited to that which makes one blameworthy: willing with particular qualities.

Secondly, one might object to some of the points made here on the following grounds: An agent who unluckily causes harm surely has reason to act and to feel differently than an agent who luckily does not. And what justifies the differing attitudes and behavior that is warranted is a difference in blameworthiness. So, for example, an agent that has negligently run over a child has reason to feel worse than an agent who was just as negligent but was lucky that no child was in the street when he careened by.

One can reply that the difference is explained by our epistemic limitations. In the case of the agent who kills the child we can be sure about the degree of negligence and carelessness. He was negligent enough to actually kill a child.

However, in the case of the agent whose negligence does not cause a death, we can still doubt the claim that he would have killed the child if she were present. Perhaps, we might think, the agent would have swerved if there had been a child in the street. That is, the tendency to think that the agents may differ with respect to blameworthiness can be explained by a difference in implicit judgments about the mental states of the respective agents.[84] But if one had complete information and was certain that the relevant mental states of the agents were similar, then one would not feel that the agents differ in blameworthiness.

Additionally, holding that agents in these kinds of cases are equal in terms of blameworthiness does not commit one to the claim that we should respond to them in the same way, legally or otherwise, or that the agents ought to feel the same.[85] One is just committed to denying that any difference in blameworthiness explains these other differences. Two golfers might make qualitatively identical swings, but due to a gust of wind after the ball is hit the shots might turn out very differently. And how the shots actually turn out can affect what the golfer ought to do next. But this is not a difference in how well the golf swing was executed (assuming that the gust of wind was not expectable). Similarly the fact that one negligent driver actually killed someone might give him reason to act differently than the agent who did not. But this is not explained by a difference in blameworthiness.

---

[84] Royzman and Kumar (2004) defend this claim on empirical grounds.

[85] See Zimmerman (2002, 562).

6.

This chapter presented an account of blameworthiness in drunk driving cases which does not appeal to the notion of tracing. This casts doubt on the claim that tracing is indispensable. This account also fits naturally with quality of will based accounts of responsibility. These accounts all hold that blameworthiness in particular and responsibility in general are solely determined by the mental states of the agent at the time of the action. And since agents can only be blameworthy for that in virtue of which they are blameworthy, agents can only be blameworthy for these mental states and hence not for the consequences of those mental states. These accounts, which do not rely on tracing, are often criticized precisely because they lack a tracing element and are thought to be unable to handle the drunk driving cases. This chapter has shown how these accounts should respond to this class of offered counterexamples. There are other putative counterexamples that these accounts must address, such as the brain manipulation cases, but those will have to be dealt with elsewhere.

CHAPTER 6

SYNCHRONIC AND DIACHRONIC RESPONSIBILITY

There are at least two different ways in which we can ascribe responsibility for an action. The first, and most commonly recognized, concerns an agent's responsibility for an action *at the time of the action*. The second concerns an agent's responsibility for an action *at some later time*. Most theorists have simply assumed that one is responsible for some past act as long as one is the same person as the author of the act. That is, that the relation that grounds responsibility through time is personal identity. I argue instead that this relation is psychological connectedness.[86] This has a number of important implications stemming from the fact that unlike personal identity this relation is scalar. In what follows I motivate and develop this distinction. I explore the way in which this distinction is relevant to the concepts of forgiveness, apology, and punishment. I then use this distinction to defend quality of will accounts of responsibility against the common manipulation cases. I argue that since the manipulations affect, by their very nature, the conditions that ground responsibility through time they do not unproblematically shed light on the conditions of responsibility at the time of the action.

Moral responsibility is taken here to be the extent to which an agent is blameworthy or praiseworthy for an action. On this conception, responsibility is a

[86] Connectedness should not be confused with continuity. Whereas continuity is not scalar and is transitive, connectedness is scalar and is not transitive. This is because continuity is the ancestral relation of connectedness; that is, continuity consists in overlapping chains of connectedness.

continuum upon which blameworthiness and praiseworthiness are poles. Thus, blameworthiness entails responsibility but responsibility does not entail blameworthiness. This sense of responsibility should be distinguished from mere causal responsibility (e.g. "carbon emissions are responsible for global warming") and from legal responsibility (e.g. "he was found responsible for misconduct").

1. The Distinction between Synchronic and Diachronic Responsibility

> Suppose that Sebastian gets drunk and drives his car on to Quincy's property. Sebastian had no intention of damaging Quincy's property. After regaining sobriety and learning of his misadventures, Sebastian, who is not yet an alcoholic, subsequently and quite deliberately returns to the local pub and proceeds to get himself roaring drunk. Though we would have excused him from moral responsibility for the damage he did on the first spree had he subsequently modified his behaviour, he did not do so and he is making the crucial past behaviour, not an out-of-character happenstance, but very much in character and hence something for which, as Aristotle would say, he may be held morally accountable (French, 1984, p.499).

French's example suggests that one's reaction to one's past act can affect one's responsibility for that act. We can imagine another version of the above story in which Sebastian, after destroying Quincy's property, promptly enrolls in a twelve step program and embarks on the road to sobriety. French's point, I take it, is that Sebastian is more blameworthy for his reckless driving in the original story than he is in the modified story. And this difference in blameworthiness is explained by Sebastian's behavior *after* the reckless driving occurs. There is something to this.

But one might resist. One might grant that Sebastian in the original story is a more blameworthy *person* than he is in the modified story. He might be

blameworthy *for his subsequent behavior* in the original version but not in the second. Yet, Sebastian's blameworthiness *for the reckless driving* does not vary across the two scenarios since the factors at the time of the action were relevantly similar; one can't undue the past. There is also something to this.

This suggests that there are two different ways in which we can ascribe responsibility. The first concerns responsibility for some action at the time of its occurrence. The second concerns responsibility for some action at some later time. I will call the former ascriptions of *synchronic responsibility* (SR) since they concern responsibility at the time of the action. Call the latter ascriptions of *diachronic responsibility* (DR) since they concern the transfer of responsibility for some action, through time, to some later time. SR concerns some agent S's responsibility at time *t* for some act X that occurs at *t*. DR concerns some agent S's responsibility at some time *t+n* for some act X that occurs at *t*. Consider Jane who, at *t*, cruelly insults her brother. In the first scenario she comes to feel a great deal of remorse about her remark and begins a program of self-improvement. At *t+1*, one year later, she is very different than she was when she issued the insult. She is no longer mean spirited, but is kind and compassionate. At this later time she cannot even consider hurting her brother's feelings to be a live option, and she offers him a sincere and heartfelt apology. In the second scenario she does not feel any remorse after the insult and does not change her ways. She does not apologize at *t+1* because she is not sorry. Jane's blameworthiness at *t* for insulting her brother is the same in the first and second scenarios. This is because her motivational structure at *t* was the same; she had the same quality of will. But

her blameworthiness at *t+1* for insulting her brother does vary across the two scenarios because an important change occurred in Jane in the first scenario and did not in the second. Her relation to that past quality of will differs in the two scenarios. In the first she distances herself from the motivations that gave rise to the insult. In the second, that motivational structure persists unchanged. Jane's blameworthiness at *t+1* for insulting her brother is mitigated in the first scenario and it is not in the second in virtue of what she is like at that later time. Jane's blameworthiness at *t* for the insult is the same in both scenarios, but her blameworthiness at *t+1* for the insult is less in the first scenario than it is in the second. The two cases are equal with respect to synchronic blameworthiness but they differ with respect to diachronic blameworthiness.[87]

One might object that the difference across the two scenarios is a difference, not in blameworthiness, but in the appropriateness of blame. That is, one might object that Jane is equally blameworthy at *t+1* in both scenarios but that it would be less appropriate to blame her in the first case than it would in the second. It has been recently argued that the conditions that make it appropriate to hold someone responsible are sensitive to a broader range of conditions than those that comprise being responsible.[88] But note that the cases that are used to support this claim focus on the person doing the blaming rather than the blameworthy agent. For example, the appropriateness of blame can, in part, depend on whether

[87] For some, a case involving a good act and a subsequent worsening of character may have more intuitive appeal.

[88] See, for example, Smith (2007).

the one doing the blaming is a victim of the harm, a bystander, or someone who has committed similar harms. In other words, whether blame is appropriate can depend on whether the evaluator has *the standing* to blame. In order for this objection to succeed, however, the factors that are supposed to explain the difference in the appropriateness of blame must be located within the agent since the cases can be constructed such that it is only internal features of the agent that vary. And it must also be clear that though these internal factors affect the appropriateness of blame they do not affect the agent's blameworthiness. It is hard for me to imagine how such a project could succeed. There is also a danger that the critic is simply using 'blameworthy' to mean 'synchronically blameworthy' and 'appropriate to blame' to mean 'diachronically blameworthy'. If this is the case then this is not an objection to my account but simply an objection to my terminology.[89]

[89] Relatedly, one might appeal to Gary Watson's (1996) distinction between responsibility as attributability and responsibility as accountability in order to explain the example. For Watson, attributability concerns whether an action can be properly ascribed to an agent. To say that an agent is to blame for an action in the attributability sense is simply to say that the agent committed the action and that the action has particular faults. Accountability, on the other hand, is conceptually tied to the notion of response. It is responsibility as accountability that we are concerned with when we are concerned with whether an agent deserves to be punished for her action. Appealing to this distinction, one might claim that at *t* the insult is attributable to Jane and that she is accountable for it. Later, at *t+1*, she is accountable only in the second scenario though the insult is still attributable to her in both scenarios. As will become evident in the next section, my claim is that responsibility as attributability varies across time. Though it is true that there is a difference in accountability at *t+1* in the two scenarios this difference is explained by a difference in attributability. This is shown when we consider cases of extreme psychological disconnectedness between the time of the act and the time of the responsibility ascription.

The conditions for synchronic and diachronic responsibility are distinct. Most of the attention in the literature has been devoted (at least implicitly) to giving conditions for synchronic responsibility. Little effort has been spent on developing the conditions for diachronic responsibility,[90] and indeed, the distinction has not (to my knowledge) been explicitly drawn. There has been much debate, for example, about whether SR is sensitive to history. Most theorists, of both compatibilist and incompatibilist stripes, believe that SR is essentially historical.[91] Quality of will (QOW) based accounts, such as Frankfurt's, deny this. This is because QOW accounts hold that responsibility is completely determined by some psychological features of the agent at the time of action.[92] Historicists claim that it matters how one came to have those psychological features; that they must have an appropriate history. The issue is whether the "snap-shot" or "current time-slice" properties[93] of an agent at the time of the action are sufficient to determine the agent's responsibility for the action at that time. This is a substantive debate. But note that it is trivially true that DR is

[90] Parfit (1984, p. 326) makes some brief but highly suggestive remarks that I attempt to develop in what follows.

[91] See, for example, Kane (1998), Fischer and Ravizza (1998), Haji (1998), and Mele (1995; 2009). Note that though these accounts do take historical considerations into account, they are not accounts of diachronic responsibility. Instead, they hold that historical considerations are relevant in determining the agent's responsibility for the action at the time of the action; that is, the agent's synchronic responsibility.

[92] See Frankfurt (1971). I also believe Strawson (1962) should be read in this way. See McKenna (2005).

[93] This language comes from Fischer and Ravizza (1998, p. 171).

sensitive to history. This is because DR concerns responsibility for something that has already happened (i.e. for some event in an agent's history). One can be responsible for some past act only if that past act occurred and this is a historical fact.

2. The Nature of Diachronic Responsibility

While there is much disagreement about the nature of SR, whether it requires libertarian free will or reasons responsiveness or alignment of one's first and second order desires, most theorists implicitly assume that DR is simply a matter of personal identity.[94] This amounts to the claim that responsibility for some action transfers through time if and only if one is the same person who committed the past act.

But the mere fact that one is the person who committed some past act does not guarantee that responsibility freely transfers. If this were so then there would be no difference in Jane's blameworthiness for the insult at *t+1* in the above example. But there is. Since it is plausible that Jane at *t* is the same person as Jane at *t+1* in both scenarios and there is a difference in blameworthiness at *t+1* between the two scenarios, this shows that DR is not simply a matter of personal

[94] David Shoemaker (unpublished manuscript) offers the following list of theorists who have accepted that "moral responsibility presupposes personal identity": Butler (1736, p. 104); DeGrazia (2005, pp. 88-89); Glannon (1998, p. 231); Haskar (1980, p. 111); Locke (1694); Madell (1981, p. 116); Parfit (1984, pp. 323-326; 1986, pp. 837-843); Reid (1785, pp. 116-117); Sider (2001, pp. 4, 143, 203-204); Schectman (1996, p. 14). I add Haji (1998), and Mele (1995).

identity. The fact that personal identity holds is not sufficient for responsibility to freely transfer through time.

Personal identity is also not necessary for responsibility to freely transfer.[95] This can be seen when one considers cases of fission. Suppose that some agent commits some crime and then undergoes fission. His two brain hemispheres are transplanted into two distinct bodies. The resulting people will both have memories of the crime, they will have the same beliefs and character traits as the pre-fission agent. It is plausible that though personal identity does not hold between either of the post-fission agents and the pre-fission agent, both post-fission agents are fully responsible for the crime.[96] "[A] malefactor could scarcely evade responsibility by contriving his own fission" (Wiggins 1976, p. 146).[97]

This suggests that DR is determined, not by personal identity, but by the relation between one's current psychology and the psychology at the time of the action. The reason that Jane's blameworthiness is mitigated in the one scenario and not in the other concerns the way her psychology at *t+1* relates to her psychology at *t*. And the reason that both post-fission agents are responsible for the crime is that their psychologies are relevantly similar to the psychology of the

---

[95] Shoemaker (unpublished manuscript) convincingly argues that responsibility does not presuppose identity.

[96] Admittedly, more would need to be said to establish this claim. I won't take up that task here since what I have to say in what follows depends on the claim that personal identity is not a sufficient condition for responsibility to freely transfer through time. It can be assumed, if one likes, both that personal identity is a necessary condition of responsibility and that it holds in the cases I discuss. Nothing hinges on this.

[97] As quoted in Parfit (1984, p. 271).

pre-fission agent. DR is thought to be a matter of personal identity, I suspect, because it often implies psychological connectedness. Psychological connectedness concerns the holding of direct psychological connections across time. One example of a direct psychological connection is that between memory and experience. A memory is directly connected to some experience if it is a memory (caused in the right way) of that very experience. But there are many other forms that psychological connectedness might take: "One such connection is that which holds between an intention and the later act in which this intention is carried out. Other such direct connections are those which hold when a belief, or a desire, or any other psychological feature, continues to be had" (Parfit, 1984; p. 205).[98] Two psychological states are connected to the degree that they are similar and causally related.

The transfer of responsibility through time is directly sensitive to psychological connectedness. In cases in which an agent commits some act X at time *t* because of some particular psychological structure then the agent's responsibility at some later time *t+n* is determined by the degree to which and the way in which that psychological structure is connected to the agent's later psychology. That is, if the agent is blameworthy for some act X to degree *d* at

[98] Parfit (1984) claims that identity matters only insofar as it implies psychological connectedness and continuity (the ancestral relation of connectedness). I follow Shoemaker (1999) who has argued that what matters is primarily connectedness and not continuity, at least in the realm of survival, anticipation, and responsibility.

time *t*, then when there is maximal connectedness between time *t* and some later time *t+n* the agent is blameworthy for X to degree *d* at *t+n*.[99]

The relevant psychological connections, as it pertains to DR, will not require global psychological connectedness. That is, the psychological connections that are relevant to DR do not include all of the agent's psychological properties, but just those that gave rise to the action; those that make up the motivational structure that brought about the act (and for that reason, are the basis of SR). If the reason Jack was rude to Jill five years ago is because he was a self-absorbed narcissist then blameworthiness can be transferred to the current time to the extent that he is still a self-absorbed narcissist. This would be true even if

[99] It may be that "maximally connected" is a threshold concept with a relatively low threshold. We may, for instance, only require a small degree of psychological connectedness in order for blameworthiness to fully transfer across time. This view is defended in Glannon (1998). He also holds that "diminished psychological connectedness does not imply diminished responsibility" (231), so long as the threshold is met. This, it seems to me, is mistaken. The mistake is to think that, with respect to some past act, an agent is either responsible for that act or not. That is, that DR is not scalar. One can be led to this view when one fails to distinguish between SR and DR. Glannon focuses on SR and it is this that leads him to deny that diminished connectedness implies diminished responsibility: "What makes a person partly or fully responsible for his behavior is not so much the strength or weakness of the connections between his earlier and later mental states, but more so his intention and his beliefs about foreseeable consequences at the time of action" (232). Were Glannon discussing the nature of SR he would be exactly right. The problem with his view is that it implies that an insignificant change in connectedness, from the threshold to just below, could be the difference between being responsible and not being responsible. The better threshold view holds that when the connectedness threshold is met DR is equal to SR, but that when the threshold is not met DR is diminished rather than eliminated. Though I leave it open here whether a threshold understanding of connectedness with respect to responsibility is best, I'm inclined to think that though we may employ a threshold understanding in practice, this is merely the best we can do given our epistemic limitations but that in fact, DR is sensitive to subtle changes in connectedness.

there is a lack of psychological connectedness in regards to something else. Jack might, for instance have cared very much about politics at the time of the insult. But he may have become disillusioned and is no longer concerned with politics at all. Psychological dissimilarity in the political realm would not, it seems, diminish his current blameworthiness for the past insult. Psychological connectedness, when judging responsibility for some past act, concerns the psychological features that led to the act. In other words, what matters for SR is the quality of will with which one acted and what matters for DR is the persistence of those qualities of will.

It is natural to think that the upper limit on one's DR for some act is set by one's SR for that act. That is, that at best DR is equal to SR because there is maximal connectedness. Yet there are cases that seem to suggest that one's DR for an act can be enhanced. Consider a case of a halfhearted insult. Suppose that Jane insults her brother at *t*. The insult was a slip of the tongue and surprised Jane. At the time she felt, to some extent, bad about it but also couldn't help feeling some joy and excitement at the jab. The insult was halfhearted. Time goes by and she no longer feels bad about it all, but is very pleased with the insult, and if she had it to do over again would have offered a crueler remark. She now, at $t+n$, wholeheartedly endorses her halfhearted insult.

When we compare this case to one in which Jane's endorsement of her insult remains halfhearted at $t+n$ I suspect that most will have the intuition that Jane in the former case is more blameworthy for the insult at $t+n$ than she is in the latter case. Why is this? At *t* Jane was motivated by particular considerations. She

was motivated to action by the prospect of hurting her brother's feelings. But she was also motivated, to some extent, to spare him harm. In the first scenario this second motivation has dissipated at *t+n* leaving only the first. Since it was the fact that she cared about hurting her brother's feelings that explained her synchronic blameworthiness for the insult, and this care has only grown in strength she is, at *t+n*, more blameworthy for the insult than she was at *t*. She more closely identifies with the bad making features of her action and is psychologically *disconnected* from the motivations that opposed it.

Enhancement cases, such as these, suggest that the valence of psychological changes matter for DR. With respect to some past motivational structure, that structure can persist as it was, it can be diminished, or it can be enhanced. When it persists one's DR remains equal to one's SR, when it is diminished one's DR becomes less than one's SR, and when it is enhanced one's DR becomes greater than one's SR.

3. Forgiveness, Apology, and Punishment

Assessments of responsibility are much more straightforward when DR more or less lines up with SR. This may explain the lack of attention paid to the distinction. In such scenarios it is only necessary to look at the agent's psychology at the time of the action. Since there has been no drastic change in psychology after the action responsibility can freely transfer.

But cases in which DR comes apart from SR are both common and troubling. They illicit in us conflict and tension. This is because it can be hard to

reconcile the differing evaluations. We seem to want and need an "all-things-considered" assessment of responsibility. But, as I have argued, DR is distinct from SR; they are different forms of evaluation.

Consider the character Ellis Boyd 'Red' Redding, played by Morgan Freeman, from the film *The Shawshank Redemption*. Red has been in the Shawshank Penitentiary since he committed an act of murder as a teenager. But the film takes place much later when Red is an aging man. His character strikes us as one of the few positive lights in the hard, violent, and often hopeless world of incarceration. He is kind and caring and the deep friendship he develops with Andy Dufresne, played by Tim Robbins, is touching. He seems like a thoroughly good person, much better than the warden, the guards, or most of the other characters in the film. But he is a murderer. He committed a horrible act. When up for a parole hearing he is asked by the committee if he feels he has been rehabilitated:

> Man #1: Your file says you've served forty years of a life sentence. You feel you've been rehabilitated?
>
> Red doesn't answer. Just stares off. Seconds tick by. The parole board exchanges glances. Somebody clears his throat.
>
> Man #1: Shall I repeat the question?
>
> Red: I heard you. Rehabilitated. Let's see now. You know, come to think of it, I have no idea what that means.
>
> Man #2: Well, it means you're ready to rejoin society as a—
>
> Red: I know what you think it means. Me, I think it's a made-up word, politician's word. A word so young fellas like you can wear a suit and tie and have a job. What do you really want to know? Am I sorry for what I did?

> Man #2: Well…are you?
>
> Red: Not a day goes by I don't feel regret, and not because I'm in here or because you think I should. I look back on myself the way I was…stupid kid who did that terrible crime…wish I could talk sense to him. Tell him how things are. But I can't. That kid's long gone, this old man is all that's left, and I have to live with that.
>
> "Rehabilitated?" That's a bullshit word, so you just go on ahead and stamp that form there, sonny, and stop wasting my damn time. Truth is, I don't give a shit (Darabont,1994, pp. 111-112).

The scene is powerful. Watching it, we can feel the frustration that Red feels being psychologically continuous but not connected with the "stupid kid who did that terrible crime." When thinking about characters like Red we can focus on the crime that occurred, the agent's SR. In many cases the agent's synchronic blameworthiness may be quite high and our emotional reactions reflect this judgment. The crime was severe, we cannot forgive, and we cannot forget. But we can, at the same time, focus on the current person before us. This person may bear little resemblance to the criminal who did the bad act. When we do this compassion comes more easily. From this perspective we can forgive and we can move on. But the fact that things look so much different from these two perspectives is difficult to reconcile.

Should people like Red be forgiven for what they have done? It is plausible that the degree to which an agent deserves to be forgiven is sensitive to the degree to which she is diachronically responsible for the prior act.[100] We can imagine cases in which an agent is so psychologically disconnected from her prior

[100] Compare Murphy: "If the wrongdoer is unrepentant, he generally does not (in my view) merit forgiveness." (2003, p. 70).

self that though there may be psychological continuity, there is no preservation of psychological content. For example, consider a case in which at time t_1 the agent is very bad and has committed some horrible act. But at some later time t_2 she is a psychological twin of, say, Mother Teresa. There are overlapping chains of psychological connectedness between t_1 and t_2 and thus there is psychological continuity between t_1 and t_2. Yet imagine that there are no direct psychological connections between t_1 and t_2; there is no preservation of psychological content between t_1 and t_2. In such cases it seems absurd to continue resenting the agent for her past acts.

DR is relevant to apology as well. It seems that to apologize for some past act is, in many cases at least, *to express* that one is psychologically disconnected from the mechanism that led to the action. When one apologizes for some past act one is expressing a normative stance towards that act; one is renouncing it. One is also often expressing that one is no longer disposed to act in that way. These are ways in which one represents oneself to be psychologically disconnected from the motivational structure that led to the action. Apology, then, often involves an expression of the relevant psychological distancing. Being sorry involves actually being relevantly psychologically disconnected. This explains why, when one is sorry and expresses this through apology, forgiveness is in order.

The distinction between SR and DR also provides a conceptual framework with which to think about the notion of punishment. Traditional retributivist approaches to punishment take the agent's SR as the central justificatory feature. That is, the extent to which an agent should be punished is determined by her

responsibility for the crime at the time of the crime; the punishment should be proportional to the crime. This approach has the advantage that it preserves the presumed "backward-looking" feature of our punishment related practices in that they are essentially *reactive*. One problem with such a strict retributivist approach, though, is illustrated by people like Red. He did commit a terrible crime, but he is so psychologically disconnected from the motivational structure that issued the crime that it can seem pointless and cruel to punish him any longer. A retributive theory that focuses on SR is open to the charge that it is blind to the way things are now. An alternative approach is to take, not the agent's SR as central, but rather what might be called her psychological flexibility. Can the criminal be rehabilitated by some mode of punishment? On this approach, the role of punishment is to encourage psychological distancing from the mechanism that issued the action. This approach is often criticized because it is exclusively "forward-looking"; it fails to account for the way in which punishment is a matter of desert. It also allows for the possibility that the rehabilitation itself is much more severe than the crime. A third approach is essentially retributivist but takes not SR but DR as central. On this view punishment is a matter of desert but the desert is determined by what the agent is like now, not what she was like at the time of the crime. This middle ground captures the attractive features of both the traditional retributive approach and the rehabilitation approach. It explains why it is pointless to punish a completely rehabilitated criminal, but it is also able to treat

punishment as essentially a matter of desert. [101] It may also explain the differing sentences we give to crimes of passion compared to crimes that were premeditated. Insofar as crimes of passion are ones in which the agents are in an abnormal psychological state of mind, it seems that the agents will be more likely to be relevantly disconnected from that state of mind after the act occurs. Crimes that are premeditated and "done with a cool head" seem to be ones in which the relevant psychological features are more likely to persist. This may be why we typically think harsher punishments are warranted in cases of the latter kind.[102] Obviously, more would need to be said to develop a retributivist account based on DR , but this suggests another avenue along which the distinction has relevance.

4. Manipulation Cases

Let us now turn to the common manipulation cases. Manipulation cases are often used to show that QOW accounts of responsibility are inadequate. That is, manipulation cases are thought to show that SR is necessarily historical and thus that QOW accounts, which are ahistorical accounts of SR, are false. Here I explore a strategy that the defender of a QOW account can employ in responding to the manipulation cases. The strategy is to argue that the intuition that the agent

[101] Consider Parfit: "Suppose that a man aged ninety, one of the few rightful holders of the Nobel Peace Prize, confesses that it was he who, at the age of twenty, injured a policeman in a drunken brawl. Though this was a serious crime, this man may not now deserve to be punished" (1984, 326).

[102] This may also explain the Aristotelian view that virtuous action must proceed from a firm and unchanging character as well as the related Humean claim that a bad action provides reason to blame the agent only if it was the result of an enduring character trait.

is not responsible in a manipulation case is explained by appeal to DR.[103] And given that we can explain the intuitive data by appeal to DR we need not hold that the manipulation cases reveal anything about the nature of SR. Consider Mele:

> *Brainwashed Beth* When Beth crawled into bed last night she was an exceptionally sweet person, as she always had been. Beth's character was such that intentionally doing anyone serious bodily harm definitely was not an option for her: her character—or collection of values—left no place for a desire to do such a thing to take root. Moreover, she was morally responsible, at least to a significant extent, for having the character she had. But Beth awakes with a desire to stalk and kill a neighbor, George. Although she had always found George unpleasant, she is very surprised by this desire. What happened is that, while Beth slept, a team of psychologists that had discovered the system of values that make Chuck [a vicious killer] tick implanted those values in Beth after erasing hers. They did this while leaving her memory intact, which helps account for her surprise. Beth reflects on her new desire. Among other things, she judges, rightly, that it is utterly in line with her system of values. She also judges that she finally sees the light about morality—that it is a system designed for and by weaklings. Upon reflection, Beth "has no reservations about" her desire to kill George and is "wholeheartedly behind it" (Frankfurt, 2002, p. 27). Furthermore, the desire is "well integrated into [her] general psychic condition" (Frankfurt, 2002, p. 27). Seeing absolutely no reason not to stalk and kill George, provided that she can get away with it, Beth devises a plan for killing him, and she executes it—and him—that afternoon…Beth is *not* morally responsible for killing George…Some readers may be inclined to believe that Beth is morally responsible for killing George. I ask such readers to add the following detail to Beth's story and to ask themselves whether it makes a difference: right after she kills George, the brainwashing is reversed (Mele 2009, pp. 464-465).

So goes one version of the manipulation case. Manipulation cases are often used as test cases for theories of responsibility. QOW theories, in particular, have often been thought to be vulnerable to these kinds of cases. These theories hold that responsibility is completely determined by some psychological features of the agent at the time of the action. The manipulation cases are then deployed as

[103] In rough slogan form: manipulations are identity undermining rather than responsibility undermining. However this can be misleading since, as I have argued, DR is not a matter of personal identity but a more subtle matter of psychological connectedness.

counterexamples to these theories. For the psychological features to which the QOW theorist appeals (note the quotes from Frankfurt in the above passage) can, it seems, be induced by artificial means. In such cases, it is thought to be intuitively clear that the agent is not responsible for the actions that result from the manipulation.[104] And this is despite the fact that the agent had the psychological features the QOW theorist holds to be sufficient for responsibility. Thus, it is claimed, the QOW account must be mistaken.

There is, undeniably, some pull to say that the agents in these cases are not responsible. The wielders of the manipulation cases claim that it is the fact that the agent has been manipulated that explains this intuition. They hold that there is an essentially historical condition on responsibility; in order for one to be responsible for an action that action must have an appropriate history (in particular, a history devoid of manipulation). But this is not the only explanation for *the not-responsible intuition*. There is an explanation of this intuition that is compatible with an ahistorical QOW approach to responsibility.

What is the source of the not-responsible intuition? There are two details of the story that seem to be doing a lot of the work. The first is that before the manipulation Beth was a nice and morally decent person. Her "character was such that intentionally doing anyone serious bodily harm definitely was not an option for her." And she was responsible for having this character. Pre-manipulation Beth was morally innocuous.

[104] The following are just a few examples of theorists who have made this claim: Fischer and Ravizza (1998), Haji (1998), Mele (2009), Pereboom (2001).

The second detail is that right after the killing the manipulation is reversed. Clearly Mele adds this feature in order to make it seem less plausible that Beth is responsible for killing.[105] So two powerful sources of the not-responsible intuition concern the moral innocuousness of Beth's psychology before and after the manipulation.[106] I suspect that the pull to say that Beth is not responsible for the killing derives from this judgment of the moral innocuousness of Beth before and after the killing, along with the judgment that she is the same person the whole time. But note that even if personal identity is held fixed across the manipulation this does not imply that the conditions of DR are also held fixed. Indeed, since DR consists in psychological connectedness it is evident that DR does not remain stable across the manipulation, since the manipulation (and its reversal) clearly involves the severing of certain psychological connections.

[105] Mele also tells "readers who are still inclined to believe that Beth is morally responsible for the killing are encouraged to replace Beth in my story with the sweetest person they know" (465). It is quite plausible that this request to replace the manipulated agent with someone we know and presumably care a great deal about (he mentions his grandmother) merely introduces bias due to special relationships. The fact that a parent finds it difficult to blame her child does not shed much light on the blameworthiness of the child. It is merely a reflection of the special relationship that the parent stands to the child. See Scanlon (2008, pp. 171-173).

[106] A third possible source of the not-responsible intuition concerns the intuition that the manipulators are responsible. This is obviously true, but there is plenty of blame to go around. The fact that one is not solely responsible for some bad act does not entail that one is not fully responsible for it (Frankfurt 1971, p. 25. fn. 10).

The manipulation itself undermines DR.[107] Pre-manipulation Beth can be thought of as $Agent_1$, who is morally innocuous. The manipulation then creates $Agent_2$. $Agent_2$ kills George and is thus is morally nocuous. She is then, through the brainwashing reversal, turned into $Agent_3$ (who is relevantly similar to $Agent_1$ and is morally innocuous). If DR were simply a matter of personal identity, and if personal identity were held fixed across the manipulations, then the moral innocuousness of $Agent_1$ and $Agent_3$ might be relevant in assessing $Agent_2$'s responsibility. But DR is not a matter of personal identity and its conditions are not held fixed across the manipulation. In the case of an extreme manipulation DR would be totally blocked. In a minor manipulation DR would be less inhibited but this would entail that there are fewer relevant psychological differences across the manipulation. And this would mean that an agent who is manipulated to kill could not have been morally innocuous before and after the manipulation: the source of the not-responsible intuition.

Suppose that in another case the manipulation is minor. The pre-manipulated agent is already disposed to kill her neighbor and the manipulation just pushes her past, what we might call, the volitional threshold; it just barely tips the scales toward killing rather than not killing. In such a case DR would be less

[107] Some theorists have noted the related claim that manipulations themselves undermine personal identity. Arpaly (2003) discusses this possibility: "But perhaps much of our tendency to look at HP (Hapless Patient, a manipulated agent) as exempt from blame comes from our response to a third kind of scenario: HP is changed into a murderer, *and then changed back*…I take it that a case involving puzzles about personal identity should not be used as a test case for theories about moral responsibility" (168). Other theorists who have considered this possibility include Fischer and Ravizza (1998, p. 235 fn. 30); Haji (1998, pp. 6-7); Mele (1995, p. 175 fn. 22); Talbert (2009), and Vargas (2006, p. 359).

inhibited across the manipulation since there are fewer relevant psychological disconnections. But I suspect that the not-responsible intuition in this case will be significantly weaker than in Mele's case in virtue of the fact that the pre-manipulated agent was already a crummy person who was disposed to kill.

Consider another story. Curt, like Beth, has killed his neighbor. And he has done so because he was, like Beth, manipulated by a team of psychologists. Many would hold that Curt is not blameworthy for killing. But suppose we learn that this psychological manipulation occurred 50 years ago, while the murder occurred today. Does the addition of this detail make an intuitive difference? It seems that when the manipulation is permanent and occurred long ago, we are less inclined to let the agent off the hook in virtue of the manipulation. Examples like these disallow the implicit assumption that guilty $Agent_2$ is turned into innocent $Agent_3$, and this explains the reluctance to excuse. In these cases we cannot ignore, as it is easy to do in cases in which the manipulation is reversed, the fact that blameworthy $Agent_2$ exists.[108]

This example also supports the claim that what matters for SR is current time-slice psychology. The intuition that the manipulated agent is not blameworthy seems to be lessened when we can rule out the possibility that the manipulation is only temporary (and for that reason that there will be an undermining of DR). That is, our intuitions might be tracking, not whether a

[108] Because I hold that manipulations can result in responsible agency this is what McKenna (2008) has called a "hard-line" reply. Indeed, all QOW theorists and, I think, all compatibilists must take a hard-line reply towards the relevant manipulation cases. This simply follows from the claim that responsible agency can arise from deterministic causal chains.

manipulation occurred, but our judgments concerning the persistence of the psychological mechanism that brought about the action.[109]

Before closing, let me sum up some of my claims regarding the manipulation cases. Responsibility needs to be indexed to time. So when thinking about responsibility in the context of a manipulation case, we need to be clear about the point at which we are assessing responsibility. The question "Is the agent responsible?" is ambiguous. It could mean any of the following:

(a) Is the pre-manipulation agent ($Agent_1$) responsible?

(b) Is the manipulated agent ($Agent_2$) responsible?

(c) Is the post-manipulation reversal agent ($Agent_3$) responsible?

All parties to the debate can agree that (a) and (c) should be answered negatively. It is question (b) that is at issue in the debate between historicists and quality of will theorists.

The wielders of the manipulation cases often implicitly appeal to the intuitiveness of negative answers to (a) and (c) in order to generate intuitive support the claim that "the agent is not responsible". They will often draw attention to the moral innocuousness of the psychology of $Agent_1$ and $Agent_3$. Note that in order for it to be plausible both that $Agent_2$ does something bad and that the pre and post-manipulation psychology is innocuous the manipulation will have to be severe.

---

[109]And there do seem to be actual cases in which the manipulation lasts long enough that we are inclined to hold the agent responsible (cases in which we can't downplay the existence of the blameworthy manipulated agent). This seemed to occur when Patty Hearst was convicted for her role in a bank heist. And it seemed that she was willing to be a part of the robbery only because she was brainwashed by her kidnappers. She was, of course, later pardoned. But this seems to be explained by the "deprogramming" that apparently occurred.

That is, if $Agent_1$ is morally unobjectionable, then since $Agent_2$ does something seriously wrong (murder, say) then it must be the case that $Agent_2$ is relevantly psychologically disconnected from $Agent_1$. Similarly for the relation between $Agent_3$ and $Agent_2$. In this way, the wielders of the manipulation cases exploit disruptions in psychological connectedness across the manipulations in order to make negative answers to (a) and (c) intuitive. But this is irrelevant since negative answers to (a) and (c) are compatible with an affirmative answer to (b), the question at issue in the debate between QOW theorists and historicists. Once we notice this, the manipulation objection is stripped of much of its intuitive clout. The question at issue is (b) and it is not clear that a negative answer to this question has any intuitive claim over a positive one.

When using thought experiments to test whether some factor F is relevant to some concept C, we will often compare cases in which F is changed and see if it amounts to a change in C. But we need to be sure that in these comparative cases we hold all other factors that might be able to affect C fixed. Otherwise we cannot be sure a change in our judgment of C is due to F. Cases involving psychological manipulation are often used as test cases for theories of SR. They are also often used to provide ground level intuitions for theory building.[110] If we are going to use examples like these to develop and evaluate theories of SR we need to be sure that these manipulations don't surreptitiously alter the conditions of DR which is then reflected in our intuitive reactions to the cases. If this is the case, then the intuitive data is contaminated. In order to legitimately draw the

[110] Fischer and Ravizza (1994), Mele (1995), and Haji (1998) for example.

conclusions that these philosophers want from these examples, we must be certain that the source of the not-responsible intuition does not concern the conditions of DR. But it is the very nature of the manipulations that they necessarily affect the conditions of DR.

5. Conclusion

Ascribing responsibility synchronically and diachronically are distinct forms of evaluation. While there are many accounts of synchronic responsibility on the table, most theorists assume that diachronic responsibility is a straightforward matter of personal identity. But this is false. Diachronic responsibility is a matter of the degree to which and the way in which one's psychology is connected to the psychology that led to the action. Manipulation cases are commonly used to raise doubts about QOW based accounts of synchronic responsibility. I've suggested that the intuitive reactions to such cases can be explained by appeal to the conditions of diachronic, rather than synchronic, responsibility. If this is right, then manipulation cases do not tell us anything interesting about synchronic responsibility and hence, cannot be used to raise problems for QOW accounts of synchronic responsibility. I've also suggested some ways in which this distinction bears on the concepts of forgiveness, apology, and punishment. It is surprising that this distinction has received so little attention given the wide scope of its relevance. Indeed, this distinction will be relevant to any area in which responsibility for something in the past is relevant.

CHAPTER 7

TYPES OF COLLECTIVE RESPONSIBILITY

The term collective responsibility is used in a variety of ways and it can refer to a number of different relations an agent, whether a collective or an individual, may bear to an action. My aim here is to distinguish these relations that fall under the sphere of collective responsibility. More specifically, I am interested in the way in which responsibility can transfer across the dimensions of space and time. I shall argue that the transfer of responsibility across these dimensions is governed by the relation of psychological connectedness.[111]

There has been a healthy (for some, decidedly unhealthy) amount of skepticism about collective responsibility. One dominant source of this skepticism concerns the fairness of collective responsibility. This has led some to go so far as to claim that "collective responsibility is barbarous" (Lewis, 1948, p.6). That is, these critics claim, it would be morally repugnant to hold individual members of a group responsible for the actions of the group solely in virtue of their group membership. One could be a member of the group yet be morally innocent, and in such a case it would be unfair to hold one responsible for the action of the group. But it is important to notice what this skepticism is about. It is about a principle of collective responsibility that says that a member of a collective is responsible for the actions of the collective solely in virtue of his membership. One can be skeptical about this principle without rejecting collective responsibility outright.

[111] As I am understanding it, an agent's moral responsibility for an act is the extent to which the agent is blameworthy or praiseworthy for the act.

We should therefore distinguish between *collective responsibility* (CR) and *individual-collective responsibility* (ICR). CR concerns the extent to which a collective is responsible for an action. ICR concerns the extent to which an individual member of a collective is responsible for an action of the collective. CR and ICR are distinguished by the subject to which a responsibility ascription is made though both are concerned with the acts of collectives. CR is concerned with the responsibility a collective bears for a collective act while ICR is concerned with the responsibility an individual member of a collective bears for a collective act. ICR distributes collective responsibility, to use a metaphor, spatially, from the collective to the individual. It is important to note that CR is conceptually prior to ICR. An account of ICR for some action must make reference to CR for that action but an account of CR for some action need not make reference to ICR for that action.[112] The concern about the fairness of collective responsibility, mentioned above, is a concern about a particular account of ICR. And the unfairness of one particular account of ICR does not entail either that all accounts of ICR or all accounts of CR violate principles of fairness.

Skepticism about collective responsibility might arise from concerns about fairness, but in a slightly different way. One might, for example, look at present day Germany and fail to find any vice that could justify holding this nation

[112] This is true even if collective responsibility is reducible to individual responsibility. This is because ICR makes essential reference to responsibility for a collective act. Whether or not a proper analysis of that collective act reduces to an analysis of individual acts, ICR must, insofar as it is concerned with collective acts, make reference to more than the individual under evaluation. It must, on a reductive approach, refer to the acts of other individuals as well.

responsible for the atrocities of the Holocaust. One might be skeptical about collective responsibility because it may seem to be insensitive to the passage of time. After all, one might ask, how could it make sense to hold a current collective responsible for some action in the past, when none of the current members were even alive at the time of action?

We should therefore distinguish between *synchronic responsibility* (SR) and *diachronic responsibility* (DR). SR concerns the extent to which an agent (whether it be an individual or a collective) is responsible at time *t* for some action that occurs at *t*. DR concerns the extent to which an agent (whether it be an individual or a collective) is responsible at some later time *t+n* for some act that occurs at *t*. It is important to note that SR is conceptually prior to DR. That is, an account of an agent's DR for some act X must make reference to her SR for X, but the reverse does not hold. My interests here primarily concern the way in which responsibility can transfer across space (from collective to individual) and time (from past to present). As such, I will not put forth an account of individual synchronic moral responsibility or of collective synchronic moral responsibility. My account will therefore be incomplete since these notions are conceptually prior to the notions with which I will be concerned. But what the present chapter lacks in comprehensiveness it makes up for in scope, since what I have to say will apply to any account of individual or collective synchronic responsibility.

Preliminaries

When one commits an action one does so for reasons. For example, if I leave my office this may be because I intend to go buy a cup of coffee. The same holds true for the actions of collectives. The reason that the US elected Barack Obama may be because of its frustration with the war in Iraq. When an individual acts she does so because of some set of beliefs, desires, intentions, behavioral dispositions, cares, and other psychological features. This is what I will call *the psychological mechanism that issued in the action*. This is simply a way of speaking about the way in which an action comes about.[113] I may have stepped on your toes because I think you are a jerk or because there was a Black Widow spider about to bite you. Though there is a clear sense in which the action is the same (stepping on your toes) it proceeds from a different psychological mechanism in these two cases. In the first this will include malevolent feelings while in the second it includes concern about your welfare. It is clear that the psychological mechanism that issues in an action is extremely relevant to moral responsibility.[114]

The same holds true for collectives. When a collective acts it does so from some particular psychological mechanism. One may find this puzzling insofar as it may seem to entail an undesirable commitment to the existence of collective

[113] My use of the notion of a psychological mechanism is similar to that of Fischer and Ravizza (1998).

[114] Indeed, in my view (which I won't defend here), it is the sole determiner of responsibility. Elsewhere I and others have referred to this as the *quality of will* with which an agent acts. I depart from that language here because 'psychological mechanism' seems a more natural description when the focus is on psychological connectedness.

psychologies, and one may find such a notion mysterious. This would be puzzling if one thought that having a psychology entailed having a brain made up of grey matter. But it only entails the existence of such psychological features as beliefs, desires, and plans. And these psychological features can be attributed to collectives.[115] One should not be skeptical of such things as mission statements and policies. Mission statements are expressions of both values and intentions, the subject of which is a collective. They express the values to which a collective is committed and declare the intention to realize those values. They are statements of the cares of the collective. Policies are norms governing the actions of the collective and are analogous to the behavioral dispositions of individuals. And we can also speak of the beliefs, desires, and other psychological states of a collective. This is not to say that in a given case it is easy to discern what, for example, the cares of a given collective are. A collective may express a commitment to certain ideals, but fail to live up to this commitment in action. It may, in a given case, be difficult to say whether the collective really cared about, say, saving the environment, or whether it was merely driven to pursue profit. But notice that the same is true of individuals as well but this does not give us reason to be skeptical of the claim that individuals can care about things.

For any action, whether it is that of an individual or of a collective, we can speak of the psychological mechanism that issued in the action. This mechanism

[115] See, for example, Bratman (1999), French (1979), Gilbert (1989), Tuomela (1989), and Velleman (1997). All that my account requires is that we can attribute these attitudes to collectives whether or not a proper account of these collective attitudes reduces to an account of the attitudes of the individual members of the collective.

will consist of beliefs, desires, cares, and other psychological features. These psychological features can persist in varying degrees. When I was a child I had the desire to become an astronaut. I no longer have this desire. To put it in another way, I am no longer *psychologically connected* to that desire. Psychological connectedness involves the holding of direct psychological connections across some dimension. One example of a direct psychological connection is that between memory and experience. A memory is directly connected to some experience if it is a memory (caused in the right way) of that very experience. But there are many other forms that psychological connectedness might take: "One such connection is that which holds between an intention and the later act in which this intention is carried out. Other such direct connections are those which hold when a belief, or a desire, or any other psychological feature, continues to be had" (Parfit, 1984; p. 205). Two psychological states are connected to the degree that they are similar and causally related.[116]

The general idea that I want to argue for here is that the distribution of responsibility whether it be from collective to individual, or from time to time, or both concerns the extent to which the agent being evaluated is psychologically connected to the mechanism that issued in the action.

[116] It should be noted that psychological connectedness is simply a way of thinking about the way in which psychological states persist over time. As such, it has no ontological commitments about the nature of such states (e.g. that an agent's desire is a concrete entity distinct from the agent).

Control

Before I begin to develop a positive account of the way the various forms of collective responsibility transfer across dimensions, I want to respond to a common objection. For some, collective responsibility is deeply unfair. This is because it amounts to holding one responsible for something that is not in one's control. This could be when an individual member of a collective is held responsible for the action of a collective, or when a collective is held responsible for some act in its history, or a combination of the two. This objection begins with the assumption that control is a necessary condition of responsibility.

One version of the objection might be put in the following way: A current member of the US is responsible for national acts of slavery in the past only if that member had control over those acts, but this is impossible. Just as an individual's responsibility for an individual act requires control, so too does responsibility for a collective act.

This objection rests on a confusion. The confusion is to think that an individual's synchronic responsibility, which plausibly may require control, is relevantly similar to an individual's responsibility for a collective action (ICR) or to a collective's responsibility for a past act (*collective diachronic responsibility* [CDR]). This is because ICR and CDR are relevantly similar, not to an individual's synchronic responsibility, but to her diachronic responsibility. Suppose that Jill insults her brother at time *t* for malicious reasons. We can distinguish between her synchronic responsibility for the insult and her diachronic responsibility for the insult. Her synchronic responsibility concerns her

responsibility at *t* for the insult which occurred at *t*. Her diachronic responsibility concerns her responsibility for the insult at some later time *t+n*. What needs to be emphasized is that only individual or collective *synchronic* responsibility may plausibly require control over the action for which one is held responsible.[117] That is, the fact that Jane is responsible at *t* for the insult may require that she had control over the insult at that time, whether through guidance control or the exercise of contra-causal freedom or control of some other kind. But it is obviously false that *diachronic* responsibility for an action requires that one has control over the action *at that later time*. That is, the fact that Jane is responsible at *t+n* for the insult which occurred at *t* obviously does not imply that Jane has control at *t+n* over the insult. It merely implies that Jane is connected to the insult in a particular way.[118]

And because ICR and CDR concern the transfer of responsibility across a dimension, they are relevantly similar to an individual's diachronic responsibility. And since diachronic responsibility for an action does not imply control over the

[117] One might object in the following way: Suppose that at t, Jones has control over whether he will go on a diet tomorrow. If he forms the intention to go on a diet tomorrow he will do so. Suppose he fails to form this intention. It may seem that he is responsible at t for failing to go on a diet tomorrow and thus, that diachronic responsibility for future acts may require control. In response I would argue that such an approach makes use of an implicit tracing principle: that one can be responsible for some later event if there is some suitable prior point that is relevantly connected to the later event. I would then respond in the way that I respond to all tracing cases. See Chapter 5.

[118] It has generally been assumed that this connection consists in personal identity. But see my "Synchronic and Diachronic Responsibility" for an argument that an individual's diachronic responsibility consists, not in personal identity, but in psychological connectedness.

action at that time, the objection to collective responsibility based on control breaks down. Control is relevant to collective responsibility in this way only if we are considering a collective's synchronic responsibility.[119] When we are considering other forms of collective responsibility control over the action is not the issue.[120] What is at issue is the way that the collective or individual under evaluation is connected to the collective act.

From Collective to Individual

Some have argued that mere membership in a collective is sufficient for inheriting the responsibilities of the collective.[121] On this view, membership is all on a par. Others have argued that members of a collective can escape responsibility for an

[119] We might, for instance, think of a proper account of collective synchronic responsibility simply as an account of the sort of control a collective must express in action in order to be responsible. French (1979) can be read as giving an account of this sort.

[120] Control may be relevant in a different way. For example, a member of a collective may be responsible for an act of the collective only if she has control over her membership. See for example, Narveson (2002).

[121] See, for example, Jaspers (1947). Arendt (1964) seems to claim that if a collective is morally responsible then so are each of its members, and then rejects collective responsibility on this basis. Raikka (1997) seems to be committed to something close to this view since he argues that a member who opposes the harmful acts of his collective is not even usually excused. Jaspers, it should be noted, distinguishes between a number of moral relations one may stand to some act. He calls the relation that all members of a collective stand to a harm committed by the collective solely in virtue of their membership metaphysical guilt. This has later gone under the heading "moral taint," see May (1992, ch. 8) and Radzik (2001).

action of the collective if they opposed the action.[122] On this view, opposition is all on a par.

But there are problems with each of these general strategies that arise from the subtlety and complexity of the phenomena under investigation. Firstly, membership is not all on a par. While it is true that there is a sense of membership in which for any given individual it is either true or it is false that that individual is a member of that collective, this is not the sense that is relevant to ascriptions of ICR. Consider the responsibility British Petroleum bears for the disaster in the Gulf. It is evident that though both the BP CEO and the oil rig worker are members of the BP collective, there is a crucial difference with respect to their responsibility for the disaster. The CEO is blameworthy to a higher degree than is the roughneck (who is probably not blameworthy at all). The sense of membership that is relevant to ICR is not a binary relation, it is scalar.

Just as there are responsibility relevant differences in membership, so too are there these differences with respect to opposition. There is a crucial difference between the man in Nazi Germany who opposes the regime by failing to report his Jewish neighbors to the SS, and the man who takes on great personal risk by hiding a family of Jews in his own home. While it is true that both of these people oppose the regime there is a clear responsibility relevant difference in their opposition. The sense of opposition that is relevant to ICR is a scalar relation.

---

[122] Raikka (1997) attributes this view to Feinberg (1968), French (1998, ch. 2), Lucas (1993), and McGray (1986).

There is a single scalar relation that explains the above considerations. This concerns the degree to which the individual member of a given collective is psychologically connected to the mechanism that issued in the collective action. The reason that the BP CEO is more blameworthy for the disaster in the Gulf than the roughneck is because he is much more psychologically connected to the mechanism that led to the disaster. The reason that the disaster occurred (I speculate) has to do with the way in which the BP collective was sensitive to particular considerations. In particular BP had institutional mechanisms in place that put the pursuit of profit over the avoidance of environmental risks in a morally objectionable way. In other words, the reason for the disaster had to do with the fact that BP, qua collective, cared too much about money and too little about reducing the risk of harm to the environment, the residents of the Gulf Coast, and even the welfare of its employees. I take it that the CEO was much more psychologically connected to these features than is the roughneck.[123] On the view I am advocating, it is the fact that the CEO is psychologically connected to the beliefs, desires, values, and so on of the collective that led to the disaster that explains his responsibility for the disaster.

Similarly, when we compare the psychological states of the German who refrains from reporting Jews to the SS but does no more to the German who does much more, what we find is a difference in the extent to which they were each psychologically (dis)connected to the mechanism that led to the Holocaust. Nazi

[123] It is, of course, possible that the roughneck is connected to the same degree as is the CEO. If this were true than the roughneck's excuse "I was only doing my job," would fail to excuse.

Germany committed the Holocaust because of a commitment to despicable ideals. The German who went to great, and personally risky, lengths to oppose the Nazi regime psychologically distances himself from those ideals to a greater extent than does the German who did less (other factors being equal). The extent to which the responsibility of a collective distributes to its members is determined by the extent to which the member is psychologically connected to the mechanism that led to the collective action.[124] If some collective C is responsible to degree *d* for some act X, then when some member of C, M, is "maximally connected" to the mechanism that led to X then M is responsible for X to degree *d*.

It is in this way that this account of ICR respects principles of individual fairness and "the separateness of persons."[125] On my account, an individual's responsibility for a collective action is not merely determined by group membership, the determiners of which may be an external matter, but by what the agent is like. Whether an individual is responsible for some collective harm is determined by the degree to which the agent is connected to the collective's flaw that led to the harm. Members who share that flaw will be more blameworthy for the harm than those members who do not.

It is worth noting that membership can correlate strongly with, and in some cases may be sufficient for, the psychological connectedness of the sort that

---

[124] This is not an entirely new idea. I take it that, for example, Feinberg's (1968) discussion of group solidarity tracks the same phenomena, as does Hill (1979). Also see May (1992, ch. 2) and Kutz (2000). What is new is the idea that a single relation, psychological connectedness, underlies the transfer of responsibility across the dimensions of space and time.

[125] Rawls (1971).

grounds ICR. Not all collectives are created equal, and membership in some may consist in being connected to the collective in a way that allows for one to be responsible for its acts. Being a citizen of the US does not consist in being connected to the mechanism that makes foreign policy decisions (one may have opposed those decisions). Being a member of the Aryan Nation does consist in being connected to the collective in such a way that one shares in the blame for its actions.[126]

Though discussions concerning the way in which responsibility can transfer between collectives and individuals have mostly focused on the way in which responsibility "trickles down" from collective to individual, it can also transfer in the reverse direction. That is, a collective may be responsible for the individual action of one of its members. Call this *collective-individual responsibility* (CIR). Suppose, for example, that at a rally a volunteer for a political campaign stomps on the head of a protester. The collective can be held responsible for that action to the extent to which it is psychologically connected to the mechanism of the individual that led to the action. Suppose, for example, that the individual committed the act from motivations of anger and hostility directed toward the group of which the protester is a member. To the extent that the collective promotes anger and hostility and uses violent language and rhetoric directed at the opposition it is connected to the individual's act and can be held

---

[126] Though this is, of course, a matter of degree. Some members may be connected to a greater degree than others. The point is that membership in some groups consists in being "connected enough" to bear blame.

responsible for it.[127] This is why when a member of a collective commits some harm the collective will often (reflexively) offer a public statement that attempts to express that they are psychologically disconnected from the act (e.g. "We do not condone these acts of violence"). This is, in effect, an attempt to evade CIR for the act.

From Past to Present

Responsibility can distribute not only across space, as it does between individuals and collectives, but also across time, as it does when some agent is currently responsible for some past harm. Recall that synchronic responsibility (SR) concerns the extent to which an agent is responsible at time *t* for an act that occurs at *t*. Diachronic responsibility (DR) concerns the extent to which an agent is responsible at some later time *t+n* for some act that occurs at *t*. In "Synchronic and Diachronic Responsibility" I explored the way that an individual agent's responsibility can vary over time. My concern here is with the way that a collective agent's responsibility can vary over time. This is what is at issue when we are wondering about the extent to which the US is now responsible for slavery.

Much of the interest in questions of the responsibility of collectives over time stems from such practical concerns as reparations for past injustice. There has been considerable effort spent on explaining how to properly identify the

[127] French (1998, Chapter 2) imagines a team of mad scientists who bear responsibility for the harm their monstrous creation has wreaked.

peoples to whom reparations are owed. Must the people to whom reparations are owed in virtue of some harm be the direct victims of that harm? Or does the claim extend to their descendants? Or, perhaps, only to those descendants that are currently harmed by the injustice?

These are difficult issues that merit careful attention, but they will not concern me presently. For successfully identifying the would-be recipients of reparations for some past injustice does not yet establish that reparations are owed. We would need to also identify the group that should pay up. The fact that some victim has been harmed by some past injustice does not entail that she is owed redress for that historical transgression, because there may be no one alive now who can be identified as the transgressor. Part of the tragedy of, for example, murder-suicides is that the one responsible cannot be *held* responsible because he is dead. This presents an immediate challenge for the case of reparations for many of the actual cases of interest involve injustices that were perpetrated by individuals who are no longer alive.

In cases in which the perpetrator of the harm can be identified, it seems clear that the responsibility to offer redress lands squarely on his (or their) shoulders. However, to set the theoretical groundwork for intergenerational reparations, some have turned to the weaker requirement that the one who owes reparations need only have benefited from the past injustice.[128] For instance, since

[128] See, for example, Boxhill (1972), Miller (2007), and Radzik (2001). Miller makes the even weaker claim that one has benefitted from membership in the group responsible for the harm, regardless of whether one benefitted from the harm itself.

white Americans have benefited from the fruits of slavery, they owe black Americans reparations. Viewed from this angle, the question of collective responsibility for past harm is a question about *whose* responsibility it is to compensate for the harm, and it is often assumed that we must assign this responsibility *to someone*.

I am interested in what I take to be a more fundamental question concerning some current collective's relation to some past wrong. This is the extent to which a collective is morally responsible for some past act, in the sense of being praiseworthy or blameworthy for that past act. For some past act for which some agent is blameworthy, there will be particular considerations in virtue of which the agent is blameworthy. The reason, for example, that the US during the Antebellum Era was blameworthy for the occurrence of slavery involves a failure to properly respect blacks. The US, as it were, failed to care enough about the welfare of all of its people. It did not care enough about not unjustly exploiting people. These are the reasons, in virtue of which, the US was blameworthy for the institution of slavery. The US, as a collective, is now blameworthy for this past wrong to the extent that those blameworthy features persist. Does the US still exhibit the racist attitudes or complicitious negligence that made slavery possible? When asking whether Germany is now responsible for the Holocaust we need to investigate whether the anti-Semitic attitudes expressed during the Nazi regime are connected to current day Germany. When investigating whether some corporation is now responsible for some past harm we need to determine whether the corporation still has the policies that resulted in the

past harm.[129] Some current collective is responsible for its past act (that is, diachronically responsible) to the extent that the collective is currently psychologically connected to the mechanism that led to the action. If some collective C is responsible to degree *d* at time *t* for some act X that occurs at *t*, then when C is "maximally connected" to the mechanism that led to X at some later time *t+n* then C is responsible for X to degree *d* at *t+n*.

Though it is no easy task to determine the extent to which a collective is psychologically connected to the mechanism that led to some past harm, this approach will, I think, yield the result that fewer collectives are responsible for past actions than do other approaches. This is partly because I am focused on responsibility in the blameworthy or praiseworthy sense, rather than the weaker notion of, for example, remedial responsibility to which David Miller (2007) appeals. But in order to establish that reparations are owed we must be able to identify a collective that is responsible in the sense that I have been concerned with. That is, to establish that reparations are now owed to some individual or group in virtue of some past harm, we must be able to identify some agent that is now blameworthy for the past harm. And this, in turn, is a question about the extent to which the agent is now psychologically connected to the mechanism that led to the past harm. Because the requirement that the current collective agent be psychologically connected to "the springs of action" of the past harm is stronger than the condition that the current collective agent has benefitted from the past harm this approach will be more conservative in its assessments of reparations.

[129] See French (1984).

Insofar as this approach yields the conclusion that, for example, the US is now not (very) blameworthy for slavery or that Germany is not now (very) blameworthy for the Holocaust, then it will turn out that issues that are commonly conceived of as questions of historical reparative justice are really questions about ahistorical distributive justice. If, for example, it is true that the US is not psychologically connected to the mechanisms that led to slavery then the US is not now blameworthy for slavery. This implies that the US does not owe reparations to current black Americans. I am not, however, claiming that the US does not owe *anything* to black Americans, just that it does not owe *reparations* (on the assumption that the current US is not appropriately connected to the acts of slavery, and therefore is not blameworthy for those acts of slavery). It is likely that black Americans are currently disadvantaged in an unjust way and the US likely has an obligation to reduce this disadvantage. But it is important to see that the reason that the US has this obligation is because there are people now who are unjustly disadvantaged. While the disadvantage has a historical explanation, it is the fact that there is current disadvantage, not that there was some past wrong, that grounds our duties toward them.

One may object to this way of thinking about collective responsibility in general, and reparations in particular, in the following way. To adapt an example of Bernard Boxhill (1972), suppose that C bought a bicycle from B. Though C took every reasonable precaution to ensure that the transaction was legal, it turns out that B had stolen the bicycle from A. Surely, it is claimed, C has a moral obligation to return the bicycle to A despite C's innocence. Similarly, regardless

of the innocence of some current collective, if that collective has benefited from some past injustice committed by it or its predecessors, the collective has a duty to make reparations for the harm.

This way of thinking about justice is pervasive. It is at the heart of the idea of "the scales of justice" and can be traced back to Aristotle.[130] The idea being that moral harmony consists in a perfectly balanced metaphorical scale. A wrong consists in moving weight from one side to the other. When B steals the bicycle from A this unbalances the scale. In order to restore harmony, it is thought, C must return the bicycle to A. When this is done the weight is again evenly distributed and all is right again. But this leaves out the crucial fact that C, who is, by stipulation, innocent has been harmed. This harm has not been made right.[131] And assuming that B has no resources (say, he has burned up the money from C), there is no adjustment of the scales that will make up for this harm. The problem with this way of thinking about justice is that, so to speak, the first law of thermodynamics does not apply to morality.[132]

Accounts of reparations that allow that a non-blameworthy collective can owe reparations to some other group is bound to combat injustice with more injustice. When C returns the bicycle to A all is not right with the universe. C has been harmed and this is unaccounted for. And suppose that A is wealthy while C

---

[130] See *Nicomachean Ethics*, book 5.

[131] I am not claiming that C has a right to the bicycle and that A does not. I am claiming that we must appeal to more than mere historical facts in order to determine where the bike ought to go.

[132] I leave it as an open question as to whether the second law does.

is poor. If this is so then the harm suffered by C returning the bicycle is much greater than the harm to A that is remedied. What this shows is that there is more of relevance than simply "righting the scales." We need to be sensitive, not only to the past, but also to the present. [133]

To reiterate: A collective is responsible at $t+n$ for its action that occurred at t to the extent that the collective is psychologically connected at $t+n$ to the mechanism that led to the action at t. Reparations require the identification of a collective that is responsible in this sense. This (probably) implies that less reparations are owed than is commonly thought. What we often think of as

---

[133] It is interesting to note that the case for reparations in the absence of diachronic responsibility may be strongest on a historical, non-patterned theory of distributive justice such as the entitlement theory of Nozick (1974):

> If the world were wholly just, the following inductive definition would exhaustively cover the subject of justice in holdings.
>
> 1. A person who acquires a holding in accordance with the principle of justice in acquisition is entitled to that holding.
> 2. A person who acquires a holding in accordance with the principle of justice in transfer, from someone else entitled to the holding, is entitled to the holding.
> 3. No one is entitled to a holding except by (repeated) applications of 1 and 2.
>
> The complete principle of distributive justice would say simply that a distribution is just if everyone is entitled to the holdings they possess under the distribution" (p. 151).

A supporter of reparations in the absence of diachronic responsibility might appeal to this entitlement theory. It could be argued that since the current distribution violates the second principle (since wealth from, e.g., slavery was improperly transferred to whites) it needs to be amended via some "principle of rectification" (Nozick 1974, p. 152). But Nozick's entitlement theory will strike many as implausible precisely because it is purely historical and is not patterned onto equality, welfare, desert, or anything else. Strong advocates of reparations and political libertarians would seem to make strange bedfellows.

questions of reparative justice for a historical harm, then, are really questions of ahistorical distributive justice.

Conclusion

There are likely those who, entrenched in liberal individualism, still find the very notion of collective responsibility troubling. They may have the inescapable feeling that collective responsibility implies that one can be responsible for things that one has not done and that this is unfair. For these people, I wish to try to make the idea more palatable.

Collective responsibility can be thought of, not as an account of the way in which some agent can be connected to an action of someone or something else, but as an account the ownership of action.[134] Individualism has too narrowly delineated the actions that can belong to an agent. If S is a member of collective C which commits action X, and S is strongly connected to the mechanism of C that led to X, then X *simply is* an action of S. In virtue of the psychological connections between S and C, S *willed* X to occur. This is why X can be attributed to S in a way that makes S open to moral appraisal in light of X.

It is a mistake to think that our physical bodies impose the limits of our agency. We act, not only as individuals, but also as collectives and this too can be an appropriate basis for moral appraisal. The extent to which one is connected to

[134] David Silver (2002) makes a similar point. I differ from him though in that I do not think that collective responsibility is *of a different kind* than is individual responsibility. The notion of responsibility is the same in both cases, what differs is the kind of agent that is involved.

the mechanism that led to an action just is the extent to which one owns the action.

I have distinguished between the responsibility of a collective qua collective, and the responsibility of individual members of the collective. I have also distinguished between responsibility at the time of action and responsibility at some later time. Combining these distinctions allows us to identify a number of different relations an agent (whether an individual or a collective) stands to an action. At the center of most moral theory is *individual synchronic responsibility* (ISR). This is the extent to which an individual is responsible at *t* for an act that occurs at *t*. But, as I have argued elsewhere, there is also *individual diachronic responsibility* (IDR). This is the extent to which an agent is responsible at *t+n* for some act that occurs at *t*. Similarly, this notion can be applied to collectives. There is both *collective synchronic responsibility* (CSR), the extent to which a collective is responsible at *t* for an act that occurs at *t*, and *collective diachronic responsibility* (CDR), the extent to which a collective is responsible at *t+n* for some act that occurs at *t*. Combining the synchronic and diachronic distinction with the relation between a collective and its members generates other responsibility relations. There is *individual-collective synchronic responsibility* (ICSR) and *individual-collective diachronic responsibility* (ICDR). These relations concern the extent to which an individual member of a collective is responsible for an action of the collective, though they are distinguished by the way that they are indexed to time. ICSR concerns the individual's responsibility at the time of the action, while ICDR concerns the individual's responsibility at

some later time. As mentioned earlier, we can also think of responsibility transferring, not from collective to individuals, but from individuals to collectives. Thus, there is *collective-individual synchronic responsibility* (CISR) and *collective-individual diachronic responsibility* (CIDR). These concern the way in which a collective can be responsible for the actions of its members. More relations can be generated since these relations are recursive, for example, that of *individual-collective-individual diachronic responsibility* (ICIDR). This is the extent to which an individual is now responsible for the prior act of another individual in virtue of their co-membership in the same collective. These various responsibility relations, aside from the conceptually prior notions of ISR and CSR, are all governed by the relation of psychological connectedness. An agent's responsibility for some act concerns the extent to which the agent is connected to the mechanism that led to the act. Psychological connectedness matters for responsibility in this way because it captures the way in which an action can belong to an agent that makes the action a proper basis for moral appraisal of the agent.

CHAPTER 8

RESPONSIBILITY AND SPEECH ACTS

P.F. Strawson famously argued that "being responsible" should be understood in terms of "holding responsible".[135] In the present chapter I take seriously the question of what is it that we are *doing* in holding another responsible. I explore the idea that responsibility should be understood on the model of speech acts.[136] The chapter argues primarily for two claims. First, that holding responsible and judging responsible share a common propositional content, the truth conditions of which simply are the conditions of being repsonsible. Second, that holding responsible involves something like a distinctive illocutionary force.[137] The various ways in which we hold responsible all involve taking that propositional content to have a particular practical significance.

1.

Sometimes when one says something this simply amounts to doing it. That is, sometimes saying it can make it so.[138] For example my utterance, "I promise to meet you tomorrow," *simply is* to promise. Similarly, when I say "I apologize for

---

[135] Strawson (1962).

[136] J.L. Austin coined the term in his influential *How to do things with Words* (1962).

[137] This idea has been suggested by theorists such as Austin, Searle, Vanderveken, and Joyce but has not, to my knowledge, been adequately developed into a theory of responsibility.

[138] See Austin (1962).

knocking over the vase," I am not describing an apology but making one. Of course, in either case I may be insincere. I may have no intention to meet you tomorrow, or I may not really be sorry. But these considerations do not make it false either that I promised or that I apologized, they just make these acts insincere.

In uttering "I promise to meet you tomorrow" I am doing something, namely promising. The illocutionary force of this utterance is one of promising. Now compare that utterance with "I warn you that I'll meet you tomorrow". This utterance shares a common propositional content with the former utterance: *I meet you tomorrow*. But despite the fact that they share propositional content they do very different things. The former promises while the second warns (or perhaps threatens). They have the same propositional content but they have different illocutionary forces.[139] Speech acts, then, can be characterized by their propositional content and their illocutionary force. We can represent this by saying that a speech act is some F(*p*) where F is some illocutionary force and *p* is some propositional content (Searle, 1969).[140]

There are many different illocutions. For example, promising, asserting, warning, and apologizing to name a few. These acts are quite diverse but there is something that they all do. They all make implications. To promise is to imply

[139] Austin (1962). Also note that the fact that some utterance has propositional content does not entail that the utterance has a truth value.

[140] This holds for the most part, but there are speech acts that lack propositional content like "Ouch!". Vanderveken labels the class of speech acts that can be understood on the F(*p*) model *elementary speech acts*.

that one has an intention to do what's been promised.[141] To assert something is to imply that one believes it. To warn is to imply danger. To apologize is to imply that one is sorry. But note that though all these illocutions make these implications, what is implied need not be the case for the illocution to occur. So long as I utter "I promise to meet you" and you hear me the promise occurs. This is true even if I have no intention of meeting you. A lack of the appropriate intention does not entail a lack of a promise, it just entails that my promise is insincere. The truth of the implications made by an illocution are part of the sincerity conditions of the illocution. These are conditions that must obtain, not for the illocution to occur, but for it to occur sincerely.[142]

To understand the nature of some illocution, then, we must understand at least three things.[143] We must investigate the nature of the propositional content of the illocution. We must understand the distinctive force of the illocution. And we should have some account of the sincerity conditions of the illocution. Here I

---

[141] In addition to *promising to* we *promise that*. Whereas a promise to carries an implication that one has the intention to do what's been promised, promising that can instead be a form of expressing one's epistemic authority with respect to some proposition (e.g. "I promise you that I was at home at the time of the shooting").

[142] Given that illocutionary acts make implications it sounds quite strange to make an illocution and then explicitly deny what's implied. That is, it would be strange to say "I promise to meet you but I have no intention of doing so" or "I apologize but I'm not really sorry". Moore may have been the first to notice this point when he noted the oddity of saying "It is raining and I do not believe it". See Moore *Principia Ethica* (1903) and Vanderveken (1990). In Austin's terms we might say that a speech act implies its felicity conditons.

[143] Though my discussion will be limited to investigating these three elements, Searle and Vanderveken identify seven elements constitutive of speech acts (1990).

will attempt to understand the notion of holding responsible on the model of speech acts. This will consist in an account of the propositional content, the illocutionary force, and the sincerity conditions. I'll proceed in that order.

2.

In this section I give an account of the propositional content of responsibility illocutions. Consider first:

> *Case 1*: Jones is on his way to work. It's raining and he is running a bit late. As he is driving he notices a motorist on the side of the road attempting to fix a flat tire. Jones, with a smile on his face, swerves to the side in order to plow through a puddle thereby soaking the stranded motorist.

I take it that most will have the intuition that Jones is clearly blameworthy for splashing the motorist. When one believes that Jones is blameworthy what is one believing?

One is believing that Jones did something that he ought not to have done.[144] He ought not to have swerved to splash the stranded motorist. He did something that, from a moral point of view, we expect him not to do. On my favored terminology, he breached a moral expectation. When one believes that some agent is blameworthy one is believing that the agent failed to meet a moral expectation.

[144] The 'ought' appealed to here is that which is discussed in Chapter 4. It is the ought that is tied to responsibility, and not the objective ought which does not take into account the agent's actual epistemic situation.

In *Responsibility and the Moral Sentiments*, R.J. Wallace develops an account of responsibility that gives this notion of expectation a central role.[145] For Wallace, an expectation can be thought of as a demand or a practical requirement, for example, "don't kill" or "it is wrong to kill" or "one ought not to kill". It is important to distinguish this sense of expectation from the epistemic sense. There is a sense of expectation that means something like "believes it more likely". For example, I may expect some of my students to cheat on the test in the sense that I think that at least one of them will actually cheat. If I were to place a bet I would wager that at least one of my students would cheat. But this is distinct from what we might call the normative sense of expectation. This is the sense in which we hold others to an expectation that they behave in some way. Despite the fact that I believe that at least one of my students will cheat, I expect them all to not cheat. That is, I hold my students to this expectation. Were one of them to violate this expectation this would warrant some negative response on my part.

Wallace distinguishes three ways in which we may affirm, in a broad sense, an expectation. First, we may internalize it. This involves being motivated to act in accordance with the expectation. For example, I may internalize an expectation concerning how much personal space is due in social situations. That is, I will actually be disposed to act in accordance with the expectation, but I may lack any evaluative beliefs regarding this expectation (Wallace, 1994, p. 44). To internalize an expectation is just to have the behavioral dispositions to act in accordance with it. Second, one may accept an expectation. Acceptance entails

[145] Wallace (1994).

internalization, according to Wallace, but it also requires a positive evaluative judgment concerning the expectation. When one accepts an expectation one is motivated to act in conformity with it and one will also believe it to be justified.[146] Thirdly, we may hold others to the expectation. Wallace argues that holding one to an expectation is a *sui generis* stance. It can not be explained by appeal to either acceptance or internalization. Rather, to hold another to an expectation is to be disposed to particular emotions or to believe those emotions are warranted in the event that the expectation is breached. The emotions implicated in holding another to an expectation are, for Wallace, the reactive attitudes. These attitudes are distinguished by their propositional content which concerns the breach of an expectation. Given this constraint, it turns out on Wallace's account that only guilt, resentment, and indignation are properly characterized as reactive attitudes for it is only these moral emotions that have this propositional object.[147] And

[146] It is not clear why we ought to think that acceptance entails internalization. There is the interesting class of cases in which one has some positive evaluative belief concerning some expectation, but has no motivation to comply. This seems to be the case with Huck Finn. He has some positive evaluative belief concerning the expectation "One ought not to help slaves escape because this amounts to stealing" in the sense that he believes it is true, yet he is not motivated to comply.

[147] Note that though it is easier to see what the reactive attitudes have in common on this narrow construal, it weakens one of Strawson's arguments for compatibilism. For the more narrow one makes the class of reactive attitudes the less plausible it becomes to think that it would be "practically inconceivable" to forswear them. That is, it does seem that it would be practically inconceivable to give up all the attitudes implicated in adult interpersonal relationships. So if one identifies the reactive attitudes with these attitudes (as Strawson does) it does seem plausible to think that we could not give them up. However if the reactive attitudes are understood narrowly, then it does seem conceivable that we could give them up. Wallace (1994).

holding responsible simply amounts to holding an agent to a moral expectation that one accepts.

Expectations can be reasonable or unreasonable. To say that an expectation is reasonable is to say that one ought to accept it, and to say that an expectation is unreasonable is to say that one ought not to accept it. Reasonable, as it is used here, should be taken to mean something like "morally justified". There is, of course, another sense of reasonable which means something like "epistemically justified". Given the distinction between the normative and epistemic senses of both 'expectation' and 'reasonable' it is possible that one can reasonably expect an agent to do something (in the epistemic sense) that it is reasonable to expect her not to do (in the normative sense). This is the case when I believe that some of my students will cheat but I hold them to a demand of academic integrity. In what follows I'll be using both 'reasonable' and 'expectation' in the normative sense.[148]

While I find Wallace's account subtle and insightful, it seems that he has only explained one side of the coin. By defining the propositional object of the reactive attitudes as a belief that an expectation has been breached, he can only explain the negative side of responsibility. Consider:

[148] Note that the debate between compatibilists and incompatibilists concerns, in my terminology, whether any expectations can be reasonable if determinism is true. I will not here enter this debate though my sympathies lie with the compatibilist.

Case 2: Smith is also running a bit late on her way to work on a rainy day. She sees a motorist on the side of the road attempting to change his tire. She continues on her way.

and

Case 3: Taylor is also on her daily commute on a rainy day and is running a bit late. She too sees a motorist attempting to change a tire. She decides to stop and help the motorist.

I take it that most will have the intuition that Smith is neither praiseworthy nor blameworthy and that Taylor is praiseworthy. Wallace does consider the question of responsibility for, what he calls, morally worthy actions. And he considers how one might extend his account to deal with cases of this kind. To hold some agent praiseworthy for some action would consist in being disposed to some positive moral emotion or to believe such emotions to be warranted in the event that agent met or exceeded an expectation one accepts.

Wallace rejects this strategy. For one, he believes that there is no moral emotion that can do the necessary work. "To hold a person responsible for a worthy action, on the other hand, does not seem presumptively connected to any positive emotions in particular" (Wallace, 1994, p. 71). He does note that we do often feel gratitude when someone does something which benefits us, but he does not think that there is a positive analogue to indignation (the third-person version of the negative reactive emotions):

> But gratitude is not called for in all cases where actions exceed the moral obligations we accept: consider the category of supererogatory acts that do not benefit us in any way. More generally, we hold people responsible for morally

> worthy acts that do not exceed the moral obligations we accept, but that merely comply with those obligations—acts such as keeping promises, telling the truth, not harming others, and so forth. In these cases it is especially clear that responsibility for worthy acts need not be connected with any distinctive sentiments (Wallace, 1994, pp. 71-72).

I believe that Wallace is wrong here. First, I think that he is simply mistaken that there is no third-person version of gratitude. Approbation seems to be a good candidate (and note, one that Strawson focused on). People are often moved to tears when they hear about the good acts, the self-sacrifice, and the extraordinary degree of compassion that some express through their actions to others. Wallace's contention that that there is no positive emotion associated with our responsibility practices gains plausibility when considering those who have merely met the expectations that we accept (e.g. keeping ordinary promises, obeying traffic laws, telling the truth, etc.). But here, he has failed to distinguish two distinct categories. It is one thing to meet an expectation, but it is another thing entirely to exceed one.

Wallace was right in thinking that the reactive attitudes share a common propositional object. And he was right to think that this object makes essential reference to expectations. But he was wrong to think that it makes essential reference to *the violation* of an expectation. There does not seem to be any reason to be this restrictive about the propositional object of the reactive attitudes. The correct account of this propositional content does make essential reference to expectations but it does not entail that an expectation has been violated. The common propositional object of the reactive attitudes and judgments of responsibility expresses that some expectation applies to the agent and expresses

that agent's relation to that expectation. But this relation does not need to be one of violation; rather it could be one of *meeting* or *exceeding* the expectation.

In Case 2 Smith simply met our moral expectations.[149] We expect people to not splash motorists for fun, but we don't expect them to always stop and help.[150] Smith's action met our expectations but did not go beyond them. Taylor, on the other hand, exceeded our expectations. Stopping to help a stranded motorist is a morally good thing, but we don't expect that people always do this (e.g. we don't react with resentment to an agent who fails to stop). Generalizing, to be blameworthy for some action is for it to be the case that that action violated a reasonable moral expectation. To be praiseworthy for some action is for it to be the case that that action went beyond or exceeded a reasonable moral expectation. And of course, there is the large (but often ignored) class of actions that do meet our reasonable moral expectations but do not go beyond them (e.g. obeying the traffic laws). This class may be characterized by the lack of any reactive attitude associated with them.

---

[149] When I refer to "our moral expectations" this should be taken to mean "the expectations that we, as a community, accept and therefore believe to be reasonable". It is a further question whether these expectations are, in fact, reasonable. Furthermore, I am not offering an account of what makes some expectation reasonable. A robust moral realist may hold that there are some Platonic moral facts that make some expectations reasonable. Various forms of anti-realism may hold that an expectation is reasonable insofar as some group or individual believes them to be reasonable. I do not want to enter this debate here. While I can remain neutral concerning whether some form of moral realism is true or not, I am committed to the denial of non-cognitivism since I hold that judgments of responsibility have an essential cognitive component (the propositional object).

[150] Perhaps we do expect them to sometimes help. In the example I am assuming that Taylor is not merely fulfilling an imperfect duty of benevolence.

To be responsible for some action, then, is to have done something to which these reasonable expectations apply. And to be a moral agent generally is just to be a person who is subject to these expectations. Some theorists use the term 'responsible' to be synonymous with 'blameworthiness'. This, it seems to me, invites confusion. For it surely seems that if one has done something praiseworthy then one is responsible. If one were not responsible for doing that praiseworthy action then it would seem that praise would be unwarranted. On my view praiseworthiness entails responsibility and blameworthiness entails responsibility, but responsibility entails neither praiseworthiness nor blameworthiness. The class of actions in which one has met but not exceeded our expectations is one we might call *mere responsibility*.

Though I won't argue the point here, the content of these reasonable expectations concerns an agent's quality of will. That is, what it is that we can reasonably expect of agents is not that they bring about this or that consequence, but that they act with particular qualities of will.[151] In my view, these qualities of will consist in acting with a reasonable degree of concern for doing what's right and avoiding what's wrong.[152]

[151] See Chapters 3 and 5.

[152] Arpaly (2003) has a similar view. Though I believe her account, as it is expressed there, is vulnerable to a serious objection. See Chapter 4.

3.

In the previous section I characterized the propositional object implicated in our various responsibility practices. This concerned an agent's relation to expectations in light of her action. This relation can be one of violating, meeting, or exceeding the expectation.

In this section I explore the question of what we are doing when we hold another responsible. Using the model of speech act theory, I will discuss the nature of the illocutionary force of holding responsible. In the previous section I offered an account of the nature of the propositional content in a speech act of holding responsible; the *p* in F(*p*). Here I will focus on the illocutionary force; the F in F(*p*).

The claim to be established is that holding responsible involves taking the agent's relation to expectations to be practically significant. This is similar to the account offered in the recent work by T.M. Scanlon.[153] For Scanlon, to be blameworthy is to have done something which impairs one's relationship with another. For example, if I reveal some secret about my friend to others in order to gain their favor, then this act of betrayal impairs my relationship with my friend. It may give my friend reason to be cautious about telling me things in confidence in the future, it may give my friend reason to think that I am less loyal than she may have originally thought, and in the extreme case it may give my friend reason

[153] Scanlon (2008).

to end our friendship. To blame another, for Scanlon, is to revise one's attitudes toward another that the impairment makes appropriate. [154]

One way in which my account differs from Scanlon's is that I want to give a comprehensive account of holding responsible. I want to capture what we are doing both when we blame and when we praise. Scanlon's account is not equipped to properly characterize praise for it is not the case that to be praiseworthy is to have done something that impairs one's relationship. Rather it would seem to enhance that relationship, to strengthen it in various ways.[155]

To hold another responsible (whether it be through praise or blame) is to adopt a policy towards that other in virtue of her relation to moral expectations. It is to give the propositional content (the agent's relation to expectations) a special practical significance or importance. For example, if I have been betrayed in some way by a friend then I will take this fact (the fact that the friend fell below a reasonable expectation) to be relevant in my dealings with the friend. The ways in which I can take this fact to be relevant are diverse and this captures the richness of our moral experience. I may hold my friend responsible by avoiding all contact with her, or I may simply shoot her a cold stare. Similarly if a friend has gone out of his way to help me, say he has given up his own kidney to save my life, then I will take this fact (that he has exceeded an expectation) to have a special

[154] Scanlon (2008, p. 128).

[155] Scanlon appears to believe that praise is a positive evaluation and, as such, is not the positive corollary to blame as he is understanding it. The positive corollary would be something like gratitude. See Scanlon (2008, pp. 151-152).

significance. I may give him a gift, express to him how much his act has meant to me, or simply be disposed to do whatever I can to return the favor.[156]

For present purposes I do not wish to argue for any particular account of the content of the policy to which one commits in holding responsible. This policy, for Wallace, would simply be a disposition to the reactive attitudes or a belief that those attitudes are warranted. But one may wish, as Scanlon does, to be more inclusive about the content of this policy. For Scanlon to hold responsible for a breaching of an expectation would just be to act in some way that reflects the impairment in the relationship that the breach of the expectation brought about. He leaves open the possibility of impairments that do not involve the reactive attitudes.

One interesting possibility is that we qua humans express the practical significance of agents' relation to expectations qua agents by way of the reactive attitudes. But we need not hold, as Wallace does, that there is a conceptual connection between holding responsible and the reactive attitudes. On this view, it is a contingent fact about human psychology that the practical significance of an agent's relation to expectations is expressed through the reactive attitudes. That is, responsibility *as we know it* essentially involves the reactive attitudes. But responsibility *simpliciter* does not essentially involve the reactive attitudes. We may wish to allow for the possibility of agents who lack the reactive attitudes yet

[156] An interesting possibility is that the fact that the friend exceeded an expectation generates in me an expectation (or perhaps, an intention or commitment) to exceed expectations in my future dealings with the friend.

who still hold responsible. Such agents, then, would express the practical significance of that propositional content by some other means.[157]

To hold another responsible is to commit oneself to a policy or to act because of and in accordance with that policy in light of an agent's relation to expectations. It is to take the agent's relation to expectations to be practically relevant. This account of holding responsible as commitment to a policy makes a related account of forgiveness, apology, and other notions possible. To blame, on this account, is to take up a policy towards an agent in virtue of that agent's breaching of an expectation. To forgive, then, is to simply to forswear that policy. When one forgives one need not revise one's judgment that the agent is responsible, that there was a breaching of an expectation. Rather, one simply takes a different stance on the significance of that fact. One no longer takes that breaching to be practically relevant in the way that is involved in blame.

Taking responsibility is the first-person version of holding responsible. When one takes responsibility for some action one is expressing that one's own relation to expectations is practically relevant for oneself and for others. Apologizing occurs when one takes responsibility for falling below an expectation. Apologizing both expresses acknowledgement that one has fallen

[157] Perhaps, for example, there are intelligent yet non-emotional creatures that alter their behavior towards others in light of that other's relation to expectations. We may wish to say that such creatures hold responsible despite a lack of emotions.

below an expectation and it involves adopting a policy to avoid a similar transgression in the future.[158]

Notice that on the account I am defending there is conceptual space for the mirror image of both forgiveness and apology. Forgiveness essentially involves taking a stance on one's breaching of an expectation. Specifically, one expresses that one no longer takes that breaching to have the practical significance involved in blame. But we may make a similar readjustment of our policy in response to someone who has exceeded our expectations. To praise, as I am understanding it, is the positive version of holding responsible. To praise an agent is to take that agent's exceeding of expectations to be practically relevant. But just as forgiveness involves no longer seeing an agent's breaching of an expectation as practically relevant, so too may we readjust our policy regarding an agent who has exceeded our expectations. We may think, for instance, that the agent has gotten enough credit already for her good deed. We may think that she has "milked it for all it's worth". And based on these considerations we may think that the agent's exceeding our expectations is no longer of practical importance.

Similarly, there may be the positive corollary of apology. Apology is a first-person version of holding responsible. To apologize is, in part, to take responsibility for breaching an expectation. But one may also take responsibility for exceeding expectations. There is no ordinary English word that refers to this

[158] As I argued in Chapter 6 this involves psychologically distancing oneself from the motivations that gave rise to the action.

phenomenon, so far as I can tell.[159] This may be because taking responsibility for a good deed would seem too self-congratulatory. And so it may be that taking responsibility for a good deed, to express that one's exceeding of an expectation is practically relevant for oneself and others, violates an expectation of humility.

To hold responsible, I've claimed, is to commit to a policy in virtue of an agent's relation to expectations. It is to make that relation important to oneself. Something has importance for one, insofar as one has particular cognitive, motivational, and emotional dispositions towards that thing.[160] That is, something is important to one when one cares about it. Holding responsible, then, essentially involves caring, in a particular way, about an agent's relation to expectations.

4.

I've claimed that to hold an agent responsible is to adopt a policy toward that agent in virtue of her relation to an expectation. This, it might be said, is the illocutionary force of holding responsible. Speech acts have sincerity conditions. To promise is to imply, among other things, that one intends to do what's been promised. But one need not actually have this intention for the promise to occur. It is possible to make an insincere promise. This raises a question. If we are modeling our account of holding responsible on the notion of speech acts, then it would seem that we should allow for cases of insincerely holding responsible. Is this a feature of our moral experience? We may distinguish between the *strong*

[159] The closest may be something like excessive pride.

[160] See Frankfurt (1982).

and *moderate* analogy with speech acts. On the strong analogy, holding responsible is an illocution just as promising is. Just as promising involves expression of a commitment to do what's been promised, holding responsible involves an expression of a commitment to a policy in light of an agent's relation to expectations. But it is commonly acknowledged that one can express a commitment to do something that one does not intend to do. Given this, the strong analogy holds that we should allow for cases in which an agent insincerely holds another responsible. The moderate analogy, on the other hand, holds that there is something to be learned from modeling our account of holding responsible on speech acts but we should not allow for the possibility of insincerely holding responsible; this would be to stretch the analogy too far.

In order to evaluate the strong and moderate analogies, we need to have some idea of what insincerely holding responsible would amount to. When one promises one makes particular implications. One implies that one has an intention to do what's been promised. A promise is insincere so long as the promiser doesn't actually have the intention that's been implied in making the illocution. On the strong version of the analogy, to insincerely hold responsible would be to imply something false. If what I've said about holding responsible is roughly right, then to hold another responsible is to imply that one takes the other's relation to expectations to have practical significance. It is to adopt a policy *purportedly* in virtue of another's relation to expectations. To hold responsible, on the strong analogy, is to imply that one takes an agent's relation to expectations to be practically relevant. To hold responsible, on the strong version, is to imply or

express that one believes the *p* in F(*p*), but one need not actually believe *p* in order for the illocution to occur.

Consider a politician who, wishing to curry favor with the populace, publicly condemns a political opponent for taking bribes. Suppose that our politician actually accepts bribes as well. Furthermore, he doesn't actually think there is anything inherently wrong with bribe taking, he just thinks that politics is a game and he is playing it better than his opponent. That is, he does not accept the expectation that his opponent breached. He does not believe the expectation to be reasonable in the normative sense. But despite this, he still expresses that he takes his opponent's actions to have practical significance. He is purporting to take his opponent's relation to expectations to have a special importance. But he does not actually think that it does because he does not accept the expectation. Such a case is, according to the strong analogy, one in which the politician insincerely holds his opponent responsible. He is actually holding the opponent responsible since his actions have a particular illocution. He commits himself to a policy that he does not believe in. His actions express that he takes the agent's relation to expectations to be practically relevant. But he fails to have particular mental states implied by the illocution (the insincere part).[161]

But one might think that this misdescribes the example. One might think that the politician is only pretending to hold responsible. He is just "going through the motions," one might be inclined to say. But because he does not have the

[161] We might think of cases of insincerely holding responsible as a species of bullshit, in Frankfurt's sense.

appropriate inner states (i.e. he doesn't believe *p*) this can't count as a case of holding responsible. To hold responsible, one might think, essentially involves a commitment to certain ideals (the expectations), a commitment that our politician lacks. On the strong analogy holding responsible just involves implying that one is committed to a policy, on the moderate analogy holding responsible requires actually committing to a policy that one believes in. The issue is whether holding responsible requires one to actually take an agent's relation to expectations as practically relevant, or to merely imply this.

There is some reason to prefer the strong analogy. One reason has to do with the reaction of the political opponent who is being blamed. For him, the overt condemnation doesn't feel like "pretend blame". It "hurts" just as much as it would if the politician had the correct inner states. Sometimes, we are more concerned with overt behavior than we are with the inner states of the person. But, on the other hand, it also seems that one can't hold responsible unless one takes morality seriously, and so there is some reason to prefer the moderate analogy.

Consider another case. Suppose there is some failed terrorist attack in Times Square. Also suppose that some militant group "takes responsibility" for the attack. And let's stipulate that this group is causally unrelated to the attack. There is a pull to say that it is true that the group "takes responsibility". They say that they do and the news reports describe the case in this way. But, one might think, the group is only "claiming responsibility", and hence they cannot actually be taking responsibility. Consider this excerpt from an actual news story:

In a purported Pakistani Taliban video that surfaced on the internet Sunday, the

> group took responsibility for the foiled attack, though Kelly said Sunday afternoon that "we have no evidence to support this claim."…"Another claim of responsibility e-mailed by an individual to a local New York news station is being investigated", Kelly said.[162]

According to the story it is true both that the group took responsibility and that they claimed responsibility. The story does say though that there is no evidence "to support this claim". What is the claim? It seems that the claim which lacks evidence is not the claim that the group took responsibility but the claim that the group *is* responsible. It doesn't seem that there is any lack of evidence that the group took responsibility. Look at the video in which the group takes responsibility. What more evidence could one want for it to be true that the group takes responsibility? So it does not seem that false claims of responsibility entail that it is false that one takes responsibility. Taking responsibility does not entail being responsible or believing that one is responsible. This, then, provides more reason to adopt the strong analogy. The terrorist group insincerely takes responsibility since they lack the appropriate belief (that they were responsible for the attack). The propositional content of the illocution concerns the speaker's relation to some expectation. This expectation would, for them, be one that they are representing that they met or exceeded and that they presumably take to be reasonable ("kill the infidels" or something along these lines), though of course we think that this expectation is not reasonable. But they don't take this proposition to be true, though they are implying that they do. This may generalize

[162] From http://www.cnn.com/2010/CRIME/05/02/times.square.closure/index.html?hpt=T1.

to other cases of holding responsible in which the speaker does not believe the appropriate proposition (some relation to some expectations).[163]

The issue, then, concerning the strong and moderate analogies concerns whether one takes the propositional content to be practically relevant. And to take the content to be practically relevant, one must believe it. There are at least two ways in which one could fail to believe that content. First, as in the case of the politician, one may have a belief concerning an agent's relation to an expectation, but one may simply fail to accept that expectation. The politician does believe that his opponent took a bribe. He just thinks that the expectation to refrain from bribery is not reasonable. The case of the terrorist group, however, is one in which the group believes the expectation to be reasonable yet they fail to have the appropriate belief concerning an agent's relation to that expectation. For them, the expectation would be, as mentioned above, something like "kill the infidels". They believe this expectation to be reasonable. But they do not believe that they were the ones who met or exceeded this expectation.

One last possibility that we should consider is that there is an asymmetry between breaching an expectation and exceeding an expectation. It may be that blame (that is, negatively holding responsible) does not require belief in the propositional content, but that praise (positively holding responsible) does require belief in the propositional content. Or perhaps the even stronger requirement that the propositional object be true. Consider a twist on the terrorist case. Suppose that rather than finding an SUV full of explosives, authorities find an SUV full of

[163] This case is also interesting since it involves a collective taking responsibility.

toys and money with instructions that they are to go to the local orphanage. And suppose that some group "takes responsibility" for leaving the SUV there but suppose that they are not in fact responsible for the good deed.[164] In such a case we may be unwilling to say that the group took responsibility so much as we would want to say that they tried (unsuccessfully) to take responsibility. Perhaps there is some condition of uptake that is part of the success conditions of this illocution. We may allow an agent to take responsibility for some bad act by just uttering the words "I hereby take responsibility" but not allow an agent to take responsibility for some good act by merely making the utterance.

What I believe the above discussion shows is that there are two distinct phenomena that we may be concerned with in the domain of holding responsible. On the one hand, we have the pure speech act which involves some overt behavior, for example, the utterance "I hereby apologize." This utterance implies that (a) one believes that one fell below some reasonable expectation and that (b) one believes this fact to have practical importance. Yet these implications need not be true in order for the apology to occur. On the other hand, we may be concerned with the inner states of the individual that make such implications true. This is what is involved in actually being sorry. When one is sorry for what one has done, one both believes that one has failed to meet a reasonable expectation and one takes this to be practically significant.

This same distinction between the speech act and the inner states implied by the speech act applies to praise and blame as well. Just we can distinguish

[164] Thanks to Cheshire Calhoun for suggesting this case.

between *apology* and *sorrow* and identify the former with the speech act and the latter with the inner states implied by the speech act, we can distinguish between *condemnation* and *blame*. To condemn involves overt behavior that implies that one has particular inner states. To blame is to actually have these inner states. Similarly, we can distinguish between *commendation* and *praise*. To commend is to go through some overt behavior that implies that one believes the expectation to have been exceeded along with the implication that this is practically relevant. Praising involves actually believing the expectation to have been exceeded and actually taking this to be of practical importance.

In normal scenarios there is a correlation between the inner states and their expression in a speech act. Blame usually gives rise to condemnation and praise to commendation. But it important to see that the two phenomena are distinct. When we speak of holding responsible we are sometimes more concerned with the inner states of the individuals but at other times we are more concerned with their overt expression. This is why both the moderate and the strong analogies had appeal. But once we clearly distinguish between the speech act and the inner states, we can see that our intuitions are not really in conflict. And what is especially important is that both the speech act and the inner states have important features in common. Both are about the practical importance of an agent's relation to expectations. This is what responsibility is about.

5.

The account of responsibility I have defended consists in two claims. The first is that responsibility is essentially about agents' relation to expectations. This relation can be one of violating, meeting, or exceeding expectations. This is the propositional content of holding responsible. I have also argued that the various ways in which we hold responsible, whether this be through praise, blame, apology, forgiveness or other modalities, all involve taking the agent's relation to expectations to be practically important. They are all ways in which we commit to a policy in virtue of one's relation to expectations. Finally, we express this commitment through overt speech acts.

REFERENCES

Anscombe, Elizabeth. 1963. *Intention*. Ithaca: Cornell University Press.

Arendt, Hannah. 1964. *Eichman in Jerusalem*. New York: Viking Press.

Aristotle. 1998. *Nicomachean Ethics.* Ross (trans.). Oxford: Oxford University Press.

Arpaly, Nomy. 2003. *Unprincipled virtue*. Oxford: Oxford University Press.

Arpaly, Nomy. 2006. *Merit, meaning, and human bondage: An essay on free will.* Princeton: Princeton University Press.

Austin, J.L. 1962. *How to do things with words*. 2nd Edition. Eds. J.O. Urmson and M. Sbisá. Cambridge, MA: Harvard University Press.

Baier, Kurt. 1998. Guilt and responsibility. In French ed. *Individvual and collective responsibility*: 93-116.

Boxhill, Bernard. 1972. The morality of reparations. *Social Theory and Practice* 2: 113-123.

Bratman, Michael. 1992. *Faces of intention: Selected essays on intentions and agency*. New York: Cambridge University Press.

Butler, Joseph. 1736. Of personal identity. In Perry ed. *Personal identity*. Berkeley: University of California Press.

Copp, David. 1997. Defending the principle of alternate possibilities: Blameworthiness and moral responsibility." *Noûs* 31: 441-456.

Darabont, Frank. 1994. *The Shawshank redemption: The shooting script*. New York: Newmarket Press.

Davidson, Donald. 1980. Agency. In *Essays on actions and events*. Oxford: Oxford University Press. 43-61.

DeGrazia, David. 2005. *Human identity and bioethics*. Cambridge: Cambridge University Press.

Domsky, Darren. 2004. There is no door: Finally solving the problem of moral luck. *The Journal of Philosophy* 101: 445-464.

Feinberg, Joel. 1968. Collective responsibility. *The Journal of Philosophy* 65: 674-688.

Fischer, John Martin and Ravizza, Mark. 1998. *Responsibility and control: A theory of moral responsibility*. New York: Cambridge University Press.

Fischer, John Martin and Tognazzini, Neal A. 2009. The truth about tracing. *Noûs* 43: 531-556.

Frankfurt, Harry. 1969. Alternate possibilities and moral responsibility. In *The importance of what we care about*. Cambridge: Cambridge University Press.

Frankfurt, Harry. 1971. Freedom of the will and the concept of a person. In *The importance of what we care about*. Cambridge: Cambridge University Press.

Frankfurt, Harry. 1982. The importance of what we care about. In *The importance of what we care about*. Cambridge: Cambridge University Press.

Frankfurt, Harry. 2002. Reply to John Martin Fischer. In Buss and Overton eds. *Contours of agency: Essays on themes from Harry Frankfurt*. Cambridge: The MIT Press.

French, Peter A. 1979. The corporation as a moral person. *American Philosophical Quarterly* 16: 207-215.

French, Peter A. 1984. A principle of responsive adjustment. *Philosophy* 59: 491-503.

French, Peter A., ed. 1998. *Individual and collective responsibility*. Rochester, VT: Schenkman.

Gilbert, Margaret. 1989. *On social facts*. New York: Routledge.

Glannon, Walter. 1998. Moral responsibility and personal identity. *American Philosophical Quarterly* 35: 231-249.

Goldman, Alvin. 1970. *A theory of human action*. Englewood Cliffs, NJ: Prentice Hall.

Haji, Ishtiyaque. 1998. *Moral appraisability: Puzzles, proposals, and perplexities.* Oxford: Oxford University Press.

Hart, H.L.A. 1968.*Punishment and responsibility: Essays in the philosophy of law*. Oxford: Clarendon Press.

Haksar, Vinit. 1980. *Equality , liberty, and perfectionism*. Oxford: Oxford University Press.

Herman, Barbara. 1993. *The practice of moral judgment*. Cambridge: Harvard University Press.

Hill, Thomas. 1979. Symbolic protest and calculated silence. *Philosophy and Public Affairs* 9: 83-102.

Jaspers, Karl. 1947. *The question of German guilt*. Trans. by E.B. Ashton. New York: The Dial Press.

Kane, Robert. 1998. *The significance of free will*. Oxford: Oxford University Press.

Kant, Immanuel. 1998. *Groundwork for the metaphysics of morals*. Mary Gregor ed. New York: Cambridge University Press.

Kutz, Christopher. 2000. *Complicity*. Cambridge: Cambridge University Press.

Lewis, H.D. 1948. Collective responsibility. *Philosophy* 24: 3-18.

Locke, John. 1689/1988. *Two treatises of government*. P. Laslett ed. Cambridge: Cambridge University Press.

Locke, John. 1694. *An essay concerning human understanding*. Partly reprinted in Perry ed. *Personal identity*. Berkeley: University of California Press.

Lucas, J.R. 1993. *Responsibility*. Oxford: Clarendon Press.

Madell, Geoffrey. 1981. *The identity of the self*. Edinburgh: Edinburgh University Press.

May, Larry. 1992. *Sharing responsibility*. Chicago: The University of Chicago Press.

McGray, Howard. 1986. Morality and collective liability. *Journal of Value Inquiry* 20: 157-165.

McKenna, Michael. 2005. Where Frankfurt and Strawson meet. *Midwest Studies in Philosophy* 29: 163-180.

McKenna, Michael. 2008. A hard-line reply to Pereboom's four-case manipulation argument. *Philosophy and Phenomenological Research*. 77: 142-159.

Mele, Alfred. 1995. *Autonomous agents*. New York: Oxford University Press.

Mele, Alfred R. 2009. Moral responsibility and history revisited. *Ethical Theory and Moral Practice* 12: 463-475.

Mill, John Stuart. 1979. *Utilitarianism*. George Sher ed. Indianapolis: Hackett Publishing Company.

Miller, David. 2007. *National responsibility and global justice*. Oxford: Oxford University Press.

Moore, G.E. 1903. *Principia Ethica*. New York: Cambridge University Press.

Murphy, Jeffrie. 2003. *Getting even: Forgiveness and its limits*. New York: Oxford University Press.

Murphy, Jeffrie and Hampton, Jean. 1988. *Forgiveness and mercy*. Cambridge: Cambridge University Press.

Nagel, Thomas. 1982. Moral luck. In D. Statman ed. *Moral luck*. Albany: State University of New York Press.

Narveson, Jan. 2002. Collective responsibility. *Journal of Ethics* 6: 179-198.

Nozick, Robert. 1974. *Anarchy, state, and utopia*. New York: Basic Books.

Parfit, Derek. 1984. *Reasons and persons*. Oxford: Oxford University Press.

Parfit, Derek. 1986. Comments. *Ethics* 96: 832-872.

Parfit, Derek. Forthcoming. *On what matters*. Oxford University Press.

Pereboom, Derk. 2001. *Living without free will*. Cambridge: Cambridge University Press.

Perry, John ed. 1975. *Personal identity*. Berkeley: University of California Press.

Radzik, Linda. 2001. Collective responsibility and duties to respond. *Social Theory and Practice* 27: 455-471.

Raikka, Juha. 1997. On dissociating oneself from collective responsibility. *Social Theory and Practice* 23: 93-108

Rawls, John. 1971. *A theory of justice*. Cambridge: Harvard University Press.

Reid, Thomas. 1785. Of Mr. Locke's account of our personal identity. In Perry ed. *Personal identity*. Berkeley: University of California Press.

Rosen, Gideon. 2008. Kleinbart the oblivious and other tales of ignorance and responsibility." *The Journal of Philosophy* 105: 591-610.

Royzman, Edward and Kumar, Rahul. 2004. Is consequential luck morally inconsequential? Empirical psychology and the reassessment of moral luck. *Ratio* 17: 329-344.

Scanlon, T.M. 2008. *Moral dimensions: Permissibility, meaning, blame*. Cambridge: Belknap Press of Harvard University Press.

Schectman, Marya. 1996. *The constitution of selves*. Ithaca: Cornell University Press.

Scott Card, Orson. 1991. *Ender's game*. New York: Tor Books.

Searle, John. 1969. *Speech acts: An essay in the philosophy of language.* Cambridge: Cambridge University Press.

Sepielli, Andrew. 2009. What to do when you don't know what to do. In R. Shafer-Landau ed. *Oxford studies in metaethics volume 4*. Oxford University Press.

Shoemaker, David. 1999. Selves and moral units. *Pacific Philosophical Quarterly* 80: 391-419.

Shoemaker, David. 2003. Caring, identification, and agency. *Ethics* 114: 88-118.

Sider, Theodore. 2001. *Four-dimensionalism*. Oxford: Oxford University Press.

Silver, David. 2002. Collective responsibility and the ownership of actions. *Public Affairs Quarterly* 16: 287-304.

Slote, Michael. 1996. Agent-based virtue ethics. *Midwest Studies in Philosophy* 20: 83-101.

Smith, Angela. 2005. Responsibility for attitudes: Activity and passivity in mental life. *Ethics* 115:236-271.

Smith, Angela M. 2007. On being responsible and holding responsible. *The Journal of Ethics* 11: 465-484.

Smith, Holly. 1983. Culpable ignorance. *The Philosophical Review* 92: 543-571.

Statman, Daniel ed. 1993. *Moral luck.* Albany: State University of New York Press.

Strawson, Peter. 1962. Freedom and Resentment. Reprinted in Watson, ed. *Free will*.

Talbert, Matthew. 2009. Implanted desires, self-formation and blame. *Journal of Ethics and Social Philosophy* 3: www.jesp.org.

Thomson, Judith Jarvis. 1971. The time of a killing. *Journal of Philosophy* 68: 115-132.

Tuomela, Raimo. 1989. Actions by collectives. *Philosophical Perspectives* 3: 471-496.

Vanderveken, D. 1990. *Meaning and speech acts, volumes I and II*. Cambridge: Cambridge University Press.

Vargas, Manuel. 2005. The trouble with tracing. *Midwest Studies in Philosophy* 29: 269-291.

Vargas, Manuel. 2006. On the importance of history for responsible agency. *Philosophical Studies* 127: 351-382.

Velleman, J.D. 1997. How to share an intention. *Philosophy and Phenomenological Research* 57: 29-50.

Wallace, R.J. 1994. *Responsibility and the moral sentiments*. Cambridge: Harvard University Press.

Watson, Gary, ed. 1982. *Free will*. New York: Oxford University Press.

Watson, Gary. 1987. Responsibility and the limits of evil: Variations on a Strawsonian theme. In *Responsibility, character, and the emotions: New essays in moral psychology*, ed. Ferdinand Shoeman. Cambridge: Cambridge University Press.

Wiggins, David. 1976. Locke, Butler and the stream of consciousness: And men as a natural kind. In Rorty ed., *The identities of persons*. Berkeley: University of California Press.

Williams, Bernard. 1981. Persons, character, and morality. In Williams, *Moral Luck*. Cambridge: Cambridge University Press.

Williams, Bernard. 1981. Moral luck. In D. Statman ed. *Moral luck*. Albany: State University of New York Press.

Zimmerman, Michael J. 1987. Luck and moral responsibility. *Ethics* 97: 374-386.

Zimmerman, Michael J. 1997. A plea for accuses. *American Philosophical Quarterly* 34: 229-243.

Zimmerman, Michael J. 2002. Taking luck seriously. *The Journal of Philosophy* 99: 553-576.

Zimmerman, Michael J. 2008. *Living with uncertainty: The moral significance of ignorance*. Cambridge: Cambridge University Press.