ABSTRACT

Title of Document: THE STRUCTURE OF RESPONSIBILITY: SYMMETRY, AGENCY, AND UNDERMINING FACTORS

Matthew David King, Ph. D, 2008

Directed By: Professor Christopher Morris, Philosophy

A THEORY OF RESPONSIBILITY ought to explain what conditions must be satisfied for an agent to be responsible for something, and whether or not ordinary agents can satisfy those conditions, given a plausible understanding of the way our world works. These goals pull against each other: the more stringent the conditions on responsibility, the harder they are to meet, and the greater the chance that we will be unable to satisfy them given a complete scientific picture of the world; the more relaxed the conditions, the easier they are to meet, but the more we may doubt their sufficiency for securing responsibility. My dissertation argues that, perhaps surprisingly, all that is required for an agent to be responsible for an action or outcome is that (1) the action was voluntary; (2) the outcome was at least foreseen; and, (3) the agent had no relevant false beliefs about the nature of what he was doing. While obviously requiring a bit of filling out and defense, these three conditions are both individually necessary and jointly sufficient for responsibility. Moreover, they are conditions that are quite easy to satisfy by ordinary agents. We should be supremely confident, therefore, that so long as we are ordinary agents, we can be responsible for the things we do.

THE STRUCTURE OF RESPONSIBILITY: SYMMETRY, AGENCY, AND UNDERMINING FACTORS

By

Matthew David King

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2008

Advisory Committee:
Professor Christopher Morris, Chair
Professor Peter Carruthers
Assistant Professor Mark Schroeder
Associate Professor Manuel Vargas
Associate Professor Karol Soltan

Foreword

Writing a dissertation (fashioning it really) is a protracted, arduous, exhilarating, and exhausting task. It has also been a deeply rewarding one. It can seem that the project is never really finished: as a culmination of one's graduate study, it is tempting to seek perfection in a dissertation. What follows is far from perfect. I still think it mostly correct, but there are still passages I'd change, phrases I'd rework, arguments I'd retool, to make it on the whole clearer, smarter, and better. But dissertations aren't finished when they're perfected. They're finished when your advisor says they are. Less cynically, more optimistically, dissertations are the first step towards working out big ideas and developing one's views, a project hopefully continued throughout one's professional career (for those of us fortunate enough to have one). It's a project I look forward to continuing.

Dedication

To my absolutely wonderful wife, Jennifer, for her undying love and support (both emotional and financial). Without her, neither the dissertation, nor much of anything in my life, would have been possible. She even helped with this dedication.

Acknowledgements

There are so many people who have played a significant role in the genesis, development, and completion of this dissertation, that I surely will miss a few. I apologize in advance, and proceed warily.

Thanks first to Sue Dwyer, who introduced me to the richness and promise of Strawson's *Freedom and Resentment*, the origin of all that follows here. Thanks to Josh Kassner, who patiently listened to my inchoate thoughts on responsibility, and more than once offered helpful insights onto the matter. Thanks especially to Bénédicte Veillet, who patiently listened to my much more fully formed thoughts on responsibility, and was quick to offer support and helpful comments. Thanks as well to all the graduate students and faculty at the University of Maryland. Each has had an influence, in small and large ways, on the project that follows.

Thanks go to my committee members. Chris Morris stepped in as official chair, and did all that was asked of him (and more). He is extraordinarily busy, and I thank him for his time and valuable comments. Peter Carruthers helped keep me from saying anything false about animals. I thank him most of all for making the time to regularly meet with me. Chapter 6 is much improved from his thoughtful remarks. Manuel Vargas is a generous mentor. From our three day philoso-thon discussing the entire dissertation over the phone, to his sharp comments at the defense, I thank him for his contributions to this dissertation and my development as a professional philosopher. I also must thank Karol Soltan, of the Department of Government and Politics at the University of Maryland, for stepping in at the last

minute, braving harsh weather and an early morning, to single-handedly make an *official* dissertation defense possible.

Last, and not at all least, I cannot thank Mark Schroeder enough. He's the smartest guy I know, and one of the sharpest philosophers around. It is impossible to tell what this dissertation would be like without his invaluable involvement. I like to remark that it is amazing what can be accomplished by a graduate student and his advisor when neither knows very much (initially) about the dissertation topic. I truly believe this dissertation is much the better for having been the result of starting from a fairly blank slate. Mark and I learned together, though I am here being presumptive, for I have little evidence that there is anything Mark doesn't know (or have a view about how to figure it out). In any case, I thank him in the deepest and most sincerest sense possible.

If I have forgotten anyone, I'm not surprised, but I am sorry. Thanks to you as well.

Table of Contents

Chapter 1: Introduction

> "[H]ow can a man be called quite free at the same moment, and with respect to the same action in which he is subject to an inevitable physical necessity? […] This is a wretched subterfuge with which some persons still let themselves be put off, and so think they have solved, with a petty word-jugglery, that difficult problem, at the solution of which centuries have laboured in vain…"
>
> – Immanuel Kant[1]

> "I'm not bad; I'm just drawn that way."
>
> – Jessica Rabbit[2]

I BELIEVE WE ARE responsible for most of what we do, in a way such that we deserve blame for the bad things we do and praise for the good things we do. And I believe the world is as science tells us: *every* event has a cause. This, surprisingly enough, is not an uncontroversial position in philosophy. But it is a respected one, with a long historical pedigree. I also think it the simplest, most natural, best-motivated, most explanatorily powerful position with the weakest theoretical commitments. In light of all this, I think it the best position to take, for it explains and secures responsibility even if the world is

[1] Kant [1788, 2002], Part 2.

[2] "Who Framed Roger Rabbit?" Touchstone Pictures, 1988.

thoroughly deterministic. But I'm getting ahead of myself. First, I want to lay out just why we should care about any position on responsibility. Second, I outline why philosophers argue about whether we are responsible for what we do or not. And third, I set out, given this background, what a philosophical theory of responsibility should do. It should explain the essential notions and core phenomena associated with responsibility, while defending the weakest set of conditions needed to do the explaining. Roughly, it should explain what is most important about responsibility in a way that is maximally compatible with a naturalistic view of the world. This dissertation is a presentation of and argument for a philosophical theory that achieves these goals.

1.1. The Importance of Responsibility – or – Why Care About a Dissertation on Responsibility?

PHILOSOPHERS, AT LEAST as early as Aristotle, have been concerned with answering the questions: (1) Do we have free will? and, (2) When is someone responsible for what they did? If a philosophical question has garnered interest for this long, two natural thoughts arise. First, it must surely be an important question for so many people to be working on answering it. And, second, how can it be that we haven't been able to answer this question for nearly 2,500 years?!

Well, while we may not have agreement among philosophers, we do have proposed answers. And we have made progress. For instance, it seems that what looked like two questions is really just one. Discussions of free will have come to focus on what it takes for someone to act *freely*, where this is understood as an important requirement for responsibility. So the big project is being able to show what it takes for people to be responsible for what they do, and the importance of free will is seen as a significant portion of that project. But this progress returns us to that first thought I mentioned: what's so important about responsibility?

The notions of freedom and responsibility, not to put too fine a point on it, lie at the center of what we care most about in life. They underwrite our views about choice and being able to direct one's own life, values enshrined in democracy and apparent in talk of political rights and expression, but are just as at home around the dinner table among teenagers arguing with parents about being able to choose for themselves. Responsibility also lies at the center of our interpersonal relationships. We see the actions of others as telling us something important about their characters, their values, what they care about. When friends do something nice for us, we treat this as expressing a concern for us because of a prior commitment to their being responsible. When a coworker snubs us, we treat it as expressing a lack of concern or an outright disrespect for us. Beyond this, our network of relationships seems to depend in important ways on treating individuals as responsible for what they do. They make things happen, because of what they believe and desire, what they value and intend. We often see people in light of their choices and commitments. What they do affects our view of who they are. Friendships and loves are built on a mutual concern for one another, one that is reinforced and exemplified by acts of kindness. Multiple transgressions against us by the same person erode a sense of concern and lead us to judge that person as holding negative attitudes about us. And such judgments depend upon a belief in responsibility.

Perhaps the most important feature of responsibility is that it is necessary for blameworthiness and praiseworthiness. We judge people all the time. We blame individuals for their transgressions against us and others. We express our outrage at perpetrators of atrocities just as we resent the guy who cuts us off in traffic. And we praise individuals for the good that they do. We hand out awards for exemplary service just as we express gratitude to those who help us. These actions and attitudes comprise a web of social practices that all depend upon individuals deserving certain types of treatment because they are responsible for what they've done. If I'm not responsible for some harm, then I cannot be blameworthy for it, and you would be mistaken to blame me. Similarly, if I'm not responsible for some benefit, then I cannot be praiseworthy, and you would be mistaken to praise me for it. Much of interpersonal life indeed is comprised of the practices of holding each other responsible, demanding that others account for their behavior, blaming and praising conduct, and offering excuses against blame and considerations that mitigate praise. All of these practices depend on seeing individuals as importantly connected to what they do and on this connection telling us something about who they are. A good person isn't just someone who does good things; it's someone who does them on purpose, out of a concern for others. A bad person isn't just someone who does bad things; it's someone who does them on purpose, who acts on a desire to hurt others (or just without concern for them). In short, the bulk of our social and moral lives depends on an important relation holding between individuals, their characters and values, and the things that they do. So, when we ask why responsibility is important, this relation gives us our answer.

It's no wonder, then, why philosophers from the beginning have been interested in giving an account of what's required to be responsible for what we do. Responsibility underwrites practices that concern much of what's deeply important about our lives: our sense of being in control of how our lives go, the value we place on interpersonal relationships like friendship and romantic love, and our practices of blaming and praising others. Given its importance, we might start to worry that we don't have a professional consensus on what the right account is. "Surely," we might say, "this isn't the sort of thing to still have unanswered!" Now that we know how important the question is, we can look at why it has been thought difficult to answer.

1.2. The Problem of Responsibility – or – Why Write a Dissertation on Responsibility?

HYPOTHETICAL EXAMPLES are often a good way to illustrate philosophical points, and the literature on responsibility is riddled with them. This dissertation is no different, so we might as well start early. Suppose that you're out to dinner at a nice Italian restaurant. You're perusing the menu. The pasta primavera looks good, but you haven't had eggplant parmesan in a while, and this particular establishment is famous for their brick-oven pizzas. If you're like most people, it seems completely up to you what you will end up ordering and eating.[3] Our futures look to us like a set of "forking paths." You could equally well choose the pasta or the eggplant or the pizza. What you choose will depend on factors like what you're in the mood for, what you ate last night, what other people at the table are ordering, etc. As this example illustrates, when faced with decisions, from

[3] I'm assuming that all three options are available in the sense that the restaurant hasn't run out of necessary ingredients and the oven isn't broken, etc.

important ones to the more mundane, we see our choices as selecting from a range of genuinely available options, and we direct the course of our lives in a significant way. If all the restaurant offered was eggplant parmesan, then besides being a very strange restaurant, your perusal of the menu would be unnecessary. It also wouldn't matter what you're in the mood for or what other people are ordering. There's only one dish to order: eggplant parmesan. It's hard in such a case to see this as being your choice. The result was inevitable given the starting conditions.

Here's the rub: there seem to be theses about the world that, if true, would make all of our choices seem like a choice of eggplant parmesan from a one-item menu. And if responsibility really requires a notion of choice that corresponds to our self-conception, then it appears that these theses are incompatible with responsibility. These theses, if true, would show that we aren't really responsible for what we do. Historically, there are a number of such theses that have been examined, but one thesis has emerged as seemingly the most dangerous for responsibility: the thesis of determinism. There are many ways to state the thesis of determinism. Here's Al Mele's helpful gloss: "[A]t any instant exactly one future is compatible with the state of the universe at that instant and the laws of nature."[4] Basically, the thesis of determinism states that the conjunction of a complete description of the universe at some instant and the laws of nature entail every subsequent truth. Every event is completely determined by the past and the laws of nature.

Determinism is a particularly dangerous thesis because of its extraordinary plausibility. Science, at least for anything bigger than quantum particles, seems thoroughly deterministic. Given a complete description of a system, the laws of nature

[4] Mele [2006], p.3.

tell us just what will happen at every instant. So it seems we cannot easily reject determinism. And if so, then the consequences seem initially dire. If determinism means that we aren't responsible, then we can only be responsible on pain of rejecting an eminently plausible scientific thesis.[5] Three options present themselves when faced with this result: one can either (1) reject determinism or (2) reject responsibility or (3) reject our self-conception of choice. These options correspond nicely to the contemporary views on responsibility. Let's examine each briefly.

Libertarians accept that responsibility is incompatible with determinism, but opt for rejecting determinism. They claim that responsibility requires our self-conception of choice, and that if determinism were true, our self-conception of choice would have to be wrong. But, they argue, determinism is false, so we can still be responsible for what we do. But rejecting determinism is a difficult path to take. First, because determinism is so plausible scientifically. Second, because rejecting it doesn't seem to secure our self-conception of choice. Science has been very successful in describing and explaining the world around us, and its achievements hardly need elaboration. I won't belabor the point here. Instead, let us suppose that the indeterminacy in quantum mechanics raises enough doubts that perhaps determinism might be false as a general thesis.

The problem for libertarians is that it isn't clear how indeterminism secures our self-conception of choice.[6] Indeterminism requires that given the same initial starting

[5] Quantum mechanics are famously indeterministic, and some have exploited this fact in their accounts (see Kane [1998]). But indeterminism doesn't fare much better in securing our self-conception of choice as I explain below.

[6] Not all libertarians resort to indeterminism for their solution. Some defend a view called "agent-causation," wherein agents can also be causes of things. But whereas deterministic laws govern events causing other events, agent-causes stand outside this deterministic framework and don't necessarily rest on indeterministic laws either. However, such views not only posit a mysterious "new" type of cause, their thesis seems even less plausible initially than purely indeterministic libertarian accounts. I'll discuss agent-causation no further.

conditions and laws of nature it was possible that you ordered a different meal then you did. At first, this looks remarkably like our self-conception of choice. But notice what the initial starting conditions include. They include, among other things, your wants and desires, your beliefs and values, etc. So indeterminacy requires that given that you wanted the eggplant parmesan most and believed you could order it, you still might not have ordered it. Suddenly, indeterminism doesn't look so promising for securing our self-conception of choice.

This isn't a knockdown argument against libertarianism, of course. Libertarians have resources to try and recapture our self-conception of choice. But they bear a burden of showing how indeterminism in the right place is enough to secure our view of choice *and* that their account is compatible with a plausible scientific worldview. This burden makes libertarianism seem, to many eyes, not a particularly attractive view.

Hard incompatibilists[7] opt for rejecting responsibility. They agree with libertarians that determinism is incompatible with our self-conception of choice, but argue that our self-conception of choice cannot be salvaged. They think our self-conception is doomed either because determinism is in fact true, or because indeterminism is no help after all. And since they hold that our self-conception of choice is necessary for responsibility, we are never responsible for what we do. Given the important role I illustrated responsibility plays, it would seem that hard incompatibilism is a position we should want to avoid. If it's true that our relationships and vital social and moral practices depend on our being responsible for things we do, then if we are not

[7] This is Derk Pereboom's label for the position. Traditionally, the rejecters of responsibility were called "hard determinists" who argued that because the thesis of determinism is true, we aren't responsible for what we do. But given the recent popularity of the second position, which includes arguments that irrespective of the truth of determinism we aren't responsible for what we do, it's simplest to bring both positions under a single heading.

responsible these things would either lose much of their value in our lives or else these practices fall apart altogether. Because of this, many hard incompatibilists additionally defend the claim that much of what we value about the practices that depend on responsibility could be salvaged without our truly being responsible for what we do.[8] However, it would seem that if we can retain all of what is important about responsibility, it would be a preferable option to hard incompatibilism.

Again, this isn't an argument against hard incompatibilism. But it's important, I think, to foreshadow the respective costs positions on the problem of responsibility incur. And it seems a reasonable aim that we should want to retain in as full a manner as possible all that is pretheoretically important to us regarding responsibility: our sense of being in control of how our lives go, the value we place on interpersonal relationships like friendship and romantic love, and our practices of blaming and praising others. Now, surely, hard incompatibilists will say they do just this, since we cannot retain responsibility in light of a plausible scientific picture of the world. Nevertheless, it seems that if we can agree on the stated goal, we should explore every theoretical avenue towards doing better than hard incompatibilism offers, if at all possible. And there is a final avenue to explore.

Compatibilists argue that responsibility and determinism are perfectly compatible. They can do this by showing that responsibility does *not* in fact depend on our self-conception of choice. Instead, they offer alternative accounts of what sorts of abilities individuals must possess in order to be responsible for what they do. They argue that the conditions necessary for responsibility can be satisfied even if determinism is true (i.e., even given the most plausible scientific picture of the world). There is, as might be

[8] For instance, see Pereboom [2001] and Sommers [2007].

expected, an enormous range of possible compatibilist views. Instead of cataloging each one, I propose to make a few comments in favor of compatibilism generally.

The promise of compatibilism is that it offers us all we could want out of a theory of responsibility. Even if the world is as science tells us it is, fear not, for we are still responsible for what we do, and so we can retain the robust notions of friendship and love, the complicated social practices of blaming and praising, our genuine sense of accomplishment at artistic creation or scholarly achievement, all of these things we deem important and valuable aspects of human life that depend on our being responsible for what we do.

Given this promise, I think compatibilism should be the default position. At least, incompatibilist positions bear the burden of arguing us out of compatibilism. If no fault can be found with a compatibilist theory, then, it should not be rejected simply because it abandons our self-conception of choice. I am tempted by this picture because I see no way to plausibly retain our self-conception if determinism is true, and I find determinism too plausible to endorse a theory that requires its falsity. I don't expect any of this to convince either libertarians or hard incompatibilists. But convincing them isn't my aim. Instead, my aim is to build the most promising compatibilist theory of responsibility, and thus, in my view, the most promising theory of responsibility.

1.3. Building a Theory of Responsibility – or – What Should a Dissertation on ResponsibilityDo?

THE PROBLEM OF RESPONSIBILITY essentially boils down to two questions. First, what conditions must individuals meet in order to be responsible for what they do? Second, are creatures such as us in a world such as ours capable of meeting those conditions?[9] These questions pull against each other. The more stringent the conditions on responsibility, the harder they are to meet, and the greater the chance that we will be unable to satisfy them given a complete scientific picture of the world. The more relaxed the conditions, the easier they are to meet, but the more we may doubt their sufficiency for securing responsibility.

This way of conceiving the dialectic is particularly helpful, since it enables a theorist to use these questions as a guide for building a theory of responsibility. This is what my dissertation does. It begins with an initial assumption: most people are responsible for many of the things they do. I think this is a plausible assumption, one that is supported by our interpersonal practices and required by many of our ordinary beliefs and conduct. Granting this assumption, one can build a theory in two "easy" steps. First, determine the essential notions and core phenomena a theory of responsibility is supposed to explain. Second, what are the weakest conditions sufficient for explaining such notions and phenomena. The aim is to provide a set of conditions maximally compatible with the world as we know it. The "best" theory would explain all that needed explaining with the conditions that are easiest to satisfy given our world.

On this model, we can even weigh the prospective merits of various theories. One that more easily secured responsibility at the cost of explaining some key notions or phenomena would bear a burden of showing why such notions were not really so

[9] Answering the first question requires a theory of what I call "local" responsibility (responsibility *for* the particular things individuals do). Answering the second requires a theory of responsible agency. A full theory of responsibility gives both answers. My approach is to answer the first question first, and then use that answer to develop an account of responsible agency. This is not, as I see it, the standard order of approach. For examples of accounts that begin (and perhaps end) with responsible agency, see Fischer and Ravizza [1998]; Frankfurt [1988]; Wolf [1990].

important, or else this fact would count against the theory. A theory that explained all the relevant data, but was unable to secure responsibility in a wide set of cases, would bear the burden of showing that explaining the phenomena was worth the cost of losing such responsibility.

1.4. My Theory of Responsibility – or – What Does My Dissertation Do?

TO CONCLUDE THIS Introduction, I want to do two things. First, try to give some sense of the historical context into which this dissertation fits. Anyone who gives a compatibilist theory of responsibility is obligated, I think, to note how it is related to the dominant compatibilist strategy, originating with Peter Strawson. His rich and insightful 1962 article sparked a fascinating program for explaining responsibility, one to which all subsequent compatibilist accounts are at least partially beholden. Second, I want to give a brief outline of the structure of the dissertation. I turn to these tasks now.

1.4.1. Strawsonian Compatiblism

THE DOMINANT COMPATIBILIST strategy for explaining responsibility is Strawsonian Compatibilism. This is a very wide label that captures many different particular accounts, all of which can stake a claim to Strawsonianism given the richness of the original article. In 1962, Peter Strawson published “Freedom and Resentment,”[10] and revolutionized the compatibilist program. The Strawsonian program consists of two prongs. The First Prong is to avoid the worries that determinism poses by explaining responsibility in terms of blameworthiness and praiseworthiness. These latter two

[10] Strawson, P. [1962].

notions he associated with our social practices of holding each other responsible, and the normative status of those practices in our lives, Strawson thought, would not be affected by the truth or falsity of determinism. No matter our metaphysical views, we would still depend on viewing others as importantly connected to those actions they did on purpose, the ones that seem to evince their attitudes towards us and the world. By translating metaphysical talk of 'being responsible' for something into the normative language of when it's appropriate to 'hold someone responsible' for something, Strawson claimed we can avoid the threats to responsibility posed by determinism. Such metaphysical theses only threaten our metaphysics, not the robust and interconnected web of normative practices that pervade our interpersonal relationships and social lives. The First Prong boils down to treating claims of being responsible as claims about when it's appropriate to blame and praise people.

The Second Prong is a method for determining what the conditions on responsibility actually are. Strawson's suggestion was to distill the conditions on responsibility out of those considerations that make it inappropriate to blame or praise individuals. As Strawson saw it, if we can rule out all the potential considerations that would render blame or praise inappropriate, then the individual in question is responsible. Those considerations that undermine responsibility, then, could give us a set of negative conditions on responsibility. So long as there is no undermining factor, the agent is responsible for the conduct in question.

In this way, Strawson considered our social practices of blaming, praising, and excusing as helpful signposts to what's necessary for responsibility. By looking to the considerations we take to make blaming or praising someone inappropriate, we highlight those considerations that render agents non-responsible for given actions. If we can rule out the presence of any such undermining factor, then the agent is responsible: we'd be liable to react to their actions in the ways distinctive of our responsibility-related practices, blaming them for bad things, and praising them for good things.

Strawson's influence cannot be underestimated. His suggestive program reoriented the compatibilist perspective on the problem responsibility. If Strawson is correct, even were we to all become convinced of determinism's truth, it wouldn't undermine our commitment to holding others responsible. Strawsonian's view was an optimistic venture, built on the importance of our social practices to our lives as agents. Of course, Strawson had and has his critics. The main worry is that there is a natural intuition that the normative status of our social practices must in some way depend on some metaphysical truths about the sorts of creatures we are. Critics are apt to claim that retaining a commitment to our social practices even in the face of evidence that we don't really have the powers or abilities we think we do is something like dogmatism. It looks to critics as though Strawsonianism simply assumes we are responsible and subsequently claims that nothing could shake our belief in this commitment. At it's best, then, Strawsonian Compatibilism is the promise of responsibility come what may; at it's worst, it is pure Pollyanna-ism.

1.4.2. The Structure of the Dissertation

I AM A COMPATIBILIST, and I think Strawson was right about a good many important things, but his program is also fundamentally flawed. This dissertation begins in Chapter 2 with the promising strategy Strawson gave us for determining the conditions on responsibility: look to the considerations that *undermine* responsibility. I examine core instances of undermining factors, like accidents, inadvertence, and mistakes. Moreover, I examine instances of undermining factors for both cases of blameworthiness and praiseworthiness. When we look across such cases, a natural observation emerges. The same considerations undermine both blameworthiness and praiseworthiness in the same way. A natural conclusion to draw is that blameworthiness and praiseworthiness must share some feature that explains why the undermining factors have this symmetrical effect. And since one can be neither blameworthy nor praiseworthy without also being responsible, a natural and plausible hypothesis is that the undermining factors work by undermining an explanatorily prior notion of responsibility shared by blameworthiness and praiseworthiness.

If this is right, it poses a severe challenge to Strawsonian Compatibilism's First Prong. Recall, the First Prong was to explain responsibility in terms of blameworthiness and praiseworthiness. But if a prior notion of responsibility is required in order to explain the symmetrical operation of the undermining factors then it appears that Strawsonian Compatibilism cannot explain compelling and core data. This is especially problematic for Strawsonianism because we arrive at the data by following the Second Prong of its own program. As a result of this difficulty, I argue we ought to reject the First Prong of Strawsonian Compatibilism in favor of pursuing the promise of the Second Prong's explanatory strategy.

I take up this task in Chapter 3, categorizing the undermining factors according to features about the agent or his action that each factor shares. Using this categorization, I pursue a suggestive hypothesis: perhaps each category of undermining factor highlights that a condition necessary for responsibility is unmet in the given case. By looking to the categories of undermining factors and the features each suggests we can actually build positive conditions on responsibility. After my categorization, we have three classes of undermining factors: those that show the action was involuntary; those that show the outcome was unintentional or unforeseen; and those that show the agent had false beliefs about the nature of his action. According to the hypothesis, these categories give us *three* positive conditions on responsibility. In order for an agent to be responsible for a particular action or outcome, it must be the case that: (1) the action was voluntary; (2) the outcome was at least foreseen; and, (3) the agent had no relevant false beliefs about the nature of his action. These are obvious simplifications of the conditions, and they require much elaboration. I conclude Chapter 3 by specifying exactly what each condition requires.

The above three conditions are, perhaps surprisingly, both individually necessary and jointly sufficient for responsibility. I defend these two claims in Chapters 4 & 5, respectively. My claim that the conditions are necessary is counterintuitive. We commonly think that negligent agents are (or at least can be) responsible for at least some of the effects of their negligent conduct. But the outcomes of negligence are often unforeseen. I show that explaining responsibility in cases of negligence is a problem for all views about responsibility, and that the standard way of explaining it, what is known in the literature as ‘tracing,’ is deeply problematic. Moreover, I give an alternative model for thinking about negligence cases that helps explain our reactions and judgments to such cases without committing to the claim that negligent agents are responsible.

Adopting my model preserves what's central about cases of negligence and defends my conditions as necessary for responsibility.

Chapter 5 examines a number of separate arguments, each of which concludes there is a further necessary condition on responsibility, one that I've omitted. I rebut these objections, in each case claiming that the arguments fail to secure their conclusions *and* that we can retain satisfactory explanations of the core phenomena involved without modifying my list of necessary conditions. In doing so, I defend my three conditions as jointly sufficient as well.

Chapter 6 brings the discussion full circle. In Chapter 2 I rejected the First Prong of the Strawsonian Compatibilist Program. The attractiveness of and motivation behind that prong was to explain the metaphysically worrisome notion of responsibility in terms of the less tricky notions of blameworthiness and praiseworthiness. Indeed, much of the attractiveness of Strawsonian Compatibilism in general lies in its promise of making the truth of determinism largely irrelevant to securing responsibility. I argue in Chapter 6 that rejecting the First Prong does not weaken my compatibilist account in the face of determinism, nor do my conditions rest on the truth of any tenuous metaphysical claims. I show that all that must be true of humans in order to be responsible is that we possess and regularly exercise three capacities, each corresponding to one of my three conditions on responsibility. And I argue that it should be uncontroversial that ordinary agents possess these capacities and routinely exercise them. We should only doubt our being responsible for what we do, therefore, to the extent that we doubt that we're ordinary agents.

Thus, my view secures responsibility for what we do and all the core phenomena associated with such a notion so long as an agent acts voluntarily to produce an outcome he foresaw and had no relevant false beliefs about what he was doing. The commitments of such a view are quite minimal and are compatible with a wide-range of plausible scientific views, including determinism. Indeed, the commitments are so weak that we should have the utmost confidence that we do satisfy them most of time. And we should therefore have the utmost confidence that we are responsible for most of what we do, in a way such that we deserve blame for the bad things we do and praise for the good things we do.

Chapter 2: The Symmetry Challenge

2.1. Introduction

I WILL BE DEFENDING a brand of compatibilism: that even in a deterministic world, where every event is entailed by the state of the world prior to that event and the laws of physics, agents can be morally responsible for what they do. I motivate my compatibilism by illustrating a serious problem that faces the most prominent compatibilist strategy currently on offer. In doing so, I aim to indirectly support an alternative account for explaining responsibility. Chapters 3-5 then examine the prospects for such an account.

A common compatibilist approach to analyzing moral responsibility is to explain it in terms of our practices of praising and blaming others. To be responsible, on this approach, is to be appropriately held responsible. The conditions on responsibility, then, are the conditions that make it appropriate to be blamed (or praised). Being responsible is simply being blameworthy or praiseworthy. Such an approach is thought to avoid difficult metaphysical commitments involved in incompatibilist accounts of responsibility, while tying the notion of responsibility intimately to practices with which we are quite familiar and about which we hold strong intuitive judgments. Call this the *Strawsonian Approach* to responsibility.[11]

This chapter argues that Strawsonian accounts[12] of responsibility fail to adequately explain a set of similarities between instances of undermined blameworthiness and undermined praiseworthiness. I claim that explaining these similarities suggests a notion of responsibility that is explanatorily prior to and significantly independent of our practices of praising and blaming. Thus, Strawsonian accounts face a significant obstacle that has not as yet been met.

[11] The classic statement of the position is P.F. Strawson [1962], "Freedom and Resentment", which focuses on the set of reactive attitudes we experience towards others in response to their conduct. A more recent and more developed account can be found in R. Jay Wallace [1994], *Responsibility and the Moral Sentiments*. Strawson's essay is a deeply rich and fertile discussion of many aspects of the free will and moral responsibility issue. Because of this complexity, many other authors consider their views "Strawsonian" while following his lead in only one of these many aspects. I of course do not take issue with every possible aspect of every plausibly Strawsonian position. My chief target in this paper is the Strawsonian strategy of explaining the notion of being responsible through our practices of praising and blaming. This aspect is crucially central to "Freedom and Resentment" and I take it forms the core of Strawson's approach to moral responsibility. The Strawsonian accounts I argue against here are unified by this strategy. For example, Jonathan Bennett claims that "someone is 'accountable' for an action…if a blame- or praise-related response to the action would not be inappropriate" (Bennett [1980], p.15). For Bennett, 'accountable' simply means 'blameworthy or praiseworthy'. Daniel Dennett suggests that understanding when individuals are responsible for what they do requires first distilling the social purpose behind *holding* others responsible (most notably by punishing those that do wrong). He states that "whatever responsibility is…unless we can tie it to some recognizable social desideratum, it will have no rational claim on our esteem" (Dennett [1984], p.163). Throughout Dennett seems chiefly concerned with the social upshots of our practices and reactive attitudes, and gives them the priority of explanation and defense. Pete Graham focuses on explaining blameworthiness in terms of the 'blame emotions' (i.e., a narrow class of reactive attitudes). As he puts it, "[the blame emotions] are the emotions the appropriate feeling of which toward someone is constitutive of that person's being blameworthy" (Graham [2005], p.5). Michael McKenna defends the Strawsonian view that "moral responsibility is constituted by a range of attitudes" (McKenna [1998], p.124). Manuel Vargas offers a revisionist version of the Strawsonian Approach to analyzing responsibility. As he puts it, the revisionist Strawsonian means "by 'S is responsible'…that there is some justified moral consideration…that entitles us to adopt towards S the stance characterized by [our] responsibility-characteristic beliefs, practices, and attitudes" (Vargas [2004], p.232). Watson expands and defends (at least portions of) Strawson's original approach in "Responsibility and the Limits of Evil." To that end, he seems to endorse the view that "[i]t is not that we hold people responsible because they *are* responsible; rather, the idea…that we are responsible is to be understood by the practice [of holding responsible]" (Watson [1987], p.258, his italics).

[12] I will refer to "the Strawsonian Approach" and "Strawsonian accounts" interchangeably. The distinctive feature of the approach, as I understand it, is that it explains 'being responsible' in terms of our practices of 'holding responsible'. Particular accounts are united by this take on the explanatory priority of our practices, but could differ in other respects.

The structure of the chapter is straightforward. Section 2 considers a partial set of data, cases in which *moral* blameworthiness and praiseworthiness are presumably undermined. I focus on this narrow set in order to draw out the problem that I think faces Strawsonian accounts. When we look at these cases, a noticeable symmetry emerges: the same considerations that undermine moral blameworthiness undermine moral praiseworthiness too. From this observation, I construct a constraint for explanatory adequacy that any account of responsibility must meet. Specifically, I claim that all accounts must be able to explain why the same factors that undermine moral blameworthiness seem to do so in the same way for moral praiseworthiness. I call this the Symmetry Challenge, and I argue that Strawsonian accounts as they are typically conceived face special difficulties in meeting it. I also provide a tentative alternative answer to the Symmetry Challenge, claiming it is both a natural and plausible response. This is to highlight the fact that the problem facing Strawsonian accounts is not a general problem for all theories of responsibility. In Section 3, I present what I take to be the best Strawsonian response to the Symmetry Challenge, drawing on the suggestion that the appropriateness of blame and praise depend on the agent in question manifesting a quality of will. Then I show why the proposed response fails by expanding on the Symmetry Challenge. I introduce the remainder of my data, widening my focus to include cases of seemingly undermined non-moral blameworthiness and non-moral praiseworthiness. In these cases, too, it is the same considerations doing the work as in the moral cases. Thus, the challenge runs even deeper than one might have initially thought, and proves a greater obstacle for the Strawsonian Approach. Section 4 provides a final revision of the Strawsonian Approach constructed to meet the challenge, and I show why it fails. As a result, Strawsonian accounts, as yet, have failed to meet the Symmetry Challenge, but their failure is instructive. In revising such an account so as to meet the challenge, success seems to depend on abandoning explanatory reliance on our practices. Thus, the Symmetry Challenge requires that explanatorily adequate theories must formulate a notion of responsibility explanatorily prior to and significantly independent of our practices. In light of my argument, I think such an analysis of responsibility is worth revisiting. I consider some final objections in Section 5.

2.2. Blameworthiness, Praiseworthiness, and Undermining Factors

2.2.1. Moral Blameworthiness and Blame-Undermining Factors

I BEGIN WITH A METHODOLOGICAL assumption: any account of responsibility ought to tell us just when an agent is responsible for something. In order to do this, I think, an account must make sense of the conditions that undermine responsibility. There are certain factors that, when present, undermine an agent's blameworthiness. Call these *blame-undermining factors*. These are also known as excuses.[13]

The Strawsonian Approach treats being responsible in terms of being appropriately held responsible. Thus, on Strawsonian accounts, being blameworthy is being appropriately blamed. So on this view, it follows that blame-undermining factors must imply that it would be inappropriate to blame an agent under the circumstances. For example, suppose Fred pokes Barney's eye. Normally, this would be an instance where

[13] It is perhaps more common to refer to such considerations as excuses or mitigating conditions. I prefer to use 'blame-undermining factors' because unlike excuses or mitigations, the language of undermining factors applies more readily to praiseworthiness. For that reason, and since my focus is on drawing particular parallels between undermined blameworthiness and praiseworthiness, I opt for the less familiar term. See also n.6, below. Moreover, I think that the term 'excuses' covers more considerations than only those that undermine *responsibility*. I take up this discussion in Chapter 2, Sec.3.

blaming Fred would be appropriate. But suppose that Fred was opening a bottle of champagne, and the cork popped, bounced off a wall and hit Barney in the eye. It seems as if Fred shot Barney in the eye, but only *accidentally*. And the fact that it was an accident seems to undermine Fred's blameworthiness for shooting Barney. Accidents are blame-undermining factors.[14]

Accidents are only one type of excuse. Suppose that Jan takes Marsha's jacket without asking. This is another case in which blame would seem appropriate. But suppose that they unknowingly wore the same jacket to the party, so Jan thought she was taking her own jacket home. It seems that Jan's blameworthiness is undermined. She meant to take her jacket, and only took Marsha's jacket by *mistake*. So, mistakes are blame-undermining factors as well.

Finally, suppose Barbie hits Ken. Here again is an intuitive case of blameworthiness. But suppose they were both riding in an elevator when Barbie suffered an epileptic fit, and in the course of her thrashing, she hit Ken. As in the cases above, Barbie's blameworthiness seems undermined too, and it seems to be the fact that her behavior was *involuntary* that does the undermining. Involuntariness, it seems, is also a blame-undermining factor.

Now we have a set of three blame-undermining factors: accident, mistake, and involuntariness. There are others, but these shall suffice for my purposes here.[15]

[14] It might seem here as if I'm relying on intuitions to make my case. But I'm not. I avoid claiming that accidents intuitively undermine responsibility. Instead, I take my examples to highlight core cases of undermining factors, cases which, I maintain, decisively and uncontroversially undermine responsibility. I claim accidents really are blame-undermining factors, not just that this is intuitively true. Since every theory must begin somewhere, my account begins with this core data. Whatever the best theory of responsibility is, therefore, I submit it would have to treat the core cases of undermining factors as undermining responsibility.

[15] At no point in this chapter do I give an exhaustive list of the undermining factors. Completeness would count in favor of such a task, but brevity and accessibility of the argument count decisively against it.

Accident, mistake, and involuntariness, when present, serve to undermine our ascriptions of blame. It follows from my methodological approach that on a Strawsonian account they do so by making it inappropriate to blame the agent. I will consider this the Strawsonian account of blame-undermining factors.

2.2.2. Moral Praiseworthiness and Praise-Undermining Factors

I HAVE ALREADY OUTLINED the very brief and basic account for a Strawsonian understanding of how undermining factors affect ascriptions of blameworthiness. When blameworthiness is undermined, it is due to some blame-undermining factor, which makes blaming the agent inappropriate. These blame-undermining factors count as considerations that mitigate our blame responses. Now, we might also think that there are considerations that mitigate our praise responses. I'll call these *praise-undermining factors*.[16] The Strawsonian Approach for praiseworthiness, then, is the analogue to the story for blameworthiness. Thus, it follows that being praiseworthy just means being appropriately praised. So, according to my methodological assumption, the Strawsonian Approach seems committed to understanding praise-undermining factors as those factors that show praise to be inappropriate under the circumstances.[17]

[16] We don't typically refer to praise-undermining factors as excuses. But to my mind, this is because excuses have become indispensable for the role they play as responses to accusations of one sort or another. We typically don't accuse others of having done something good, and so praise-undermining factors are not needed for defenses in such cases. Nonetheless, we do often refer to praise-undermining factors to show certain responses to others as inappropriate. This section highlights some examples of when this is the case. They could also be called praise-mitigating factors, or even "*praise*cuses".

[17] There could be two ways we think praise inappropriate. First, we might not think anything good has really come about. Second, we might agree that good has come about, but think that nevertheless the agent isn't praiseworthy on account of the outcome. I intend for the following examples to highlight the second of these thoughts. Should an example prompt the first thought, the reader is invited to construct a parallel case where good does come about.

Suppose that Bruce saves a child’s life. Certainly, this is normally a meritorious deed. But suppose that Bruce was fishing off the pier, when his hook caught something. As he reels in what he thinks is a large fish, it turns out he hooked a drowning boy by his jacket. It surely seems as if praise would be inappropriate in this case. Bruce didn’t mean to save the child; he only did so accidentally. And the fact that it was an accident seems to undermine his praiseworthiness. Accidents are praise-undermining factors.

Imagine that Diana thinks she’s adding sugar to a customer’s coffee when she’s actually adding his heart medicine, without which he’ll die very shortly. Diana saves the man’s life, and normally this would be a laudable deed, but under the circumstances it doesn’t seem to be appropriate to praise her. After all, she only saves his life by mistake. So, it seems as if the mistake serves to undermine her praiseworthiness. Mistake is also a praise-undermining factor.

Finally, suppose Clark saves a man’s life in the elevator. This would indeed normally be a commendable act. But suppose Clark suffered a seizure and in his flailing hit the man below the solar-plexus. As it turns out, the man had just begun to choke, and Clark’s punch dislodged the culprit. Clark saved the man’s life, but he did so involuntarily. While we might be amazed at the “lucky” circumstances, it seems as if praising Clark in this case would be inappropriate. It would be inappropriate because, while he saved the man’s life, he did so involuntarily. As above, involuntariness, it seems, is a praise-undermining factor as well.

Now we have a set of three praise-undermining factors: accident, mistake, and involuntariness. There are still others, but these will do for now. When present, accident, mistake, and involuntariness serve to undermine our ascriptions of praise. According to the Strawsonian, they do so by making it inappropriate to praise the agent. I will consider this the Strawsonian account of praise-undermining factors.

There is a natural observation to make at this point. Our set of blame-undermining factors is *identical* to our set of praise-undermining factors.[18] The very same factors that make blame inappropriate also make praise inappropriate. We now have a challenge we can present to any theory of responsibility:

> **The Symmetry Challenge:** Any account of the undermining factors must explain why the same factors undermine both blameworthiness and praiseworthiness, and do so in the same way.

Failure to acknowledge the challenge commits one to denying that both blame and praise are undermined in the above cases. And since I've assumed that any theory of responsibility must explain just when and why agents are or are not responsible in particular cases, failing to answer the challenge constitutes a failure to explain compelling theory-neutral phenomena regarding responsibility. I will argue shortly that Strawsonian accounts have not yet met the challenge.

2.2.3. A Simple Solution

IT IS WORTH NOTING that the Symmetry Challenge is not a problem for every theory of responsibility. For instance, the obvious conclusion to draw from the evidence above is that undermining factors affect something cases of blameworthiness and praiseworthiness

[18] I have used 'accident', 'mistake', and 'involuntariness' as representative examples of undermining factors. As noted above, I do not mean them to be an exhaustive list. I leave it to the reader to run parallel test cases with other common factors, such as 'inadvertence' and 'ignorance'.

share. And given the plausible assumption that being morally responsible for something is a necessary condition for one's being blameworthy or praiseworthy for it, a natural inference to draw is that the shared component of such cases involves an explanatorily prior and independent notion of responsibility.[19] I'll call this the *Simple Solution* to the Symmetry Challenge.[20] Blameworthiness, on this account, is explained in terms of being responsible for something morally bad. Praiseworthiness is explained in terms of being responsible for something morally good. An extremely appealing explanation of how these factors undermine blameworthiness and praiseworthiness, then, is that they do so in virtue of showing that the agent was not responsible for the thing in question, where the conditions on responsibility are in some suitable way independent of and met prior to those for blameworthiness or praiseworthiness. More needs to be said to fill such an account out, but it meets the Symmetry Challenge in a straightforward and plausible way.

2.3. The Strawsonian Reply

BUT CAN THE STRAWSONIAN Approach answer the challenge? It could, if it could show that its accounts of blameworthiness and praiseworthiness share some feature that an answer could exploit. That would show how the same considerations could affect both

[19] By 'independent' here, I mean independent of the evaluative component of properties of blameworthiness or its positive analogue praiseworthiness. Any view that takes the conditions on moral responsibility to be explanatorily prior (or more fundamental) than those on praiseworthiness and blameworthiness would qualify. Examples seem to include (here I consider compatibilists only) Frankfurt's view, where one is responsible so long as one has a second-order desire to have the first-order desire that prompted the action, or possibly Fischer and Ravizza's view, where one is responsible so long as (roughly) one has guidance control over the action. These are obvious simplifications of both views, but in each case, the conditions on responsibility seem to be independent of a characterization of blameworthiness or praiseworthiness. For a more detailed presentation of these views, see Frankfurt [1988]; and Fischer and Ravizza [1998].

[20] Manuel Vargas has a similar answer, what he calls his "agent-based account". See Vargas [ms 1].

blameworthiness and praiseworthiness. So, recall how the Strawsonian Approach distinctively understands blameworthiness and praiseworthiness. To be blameworthy is to be appropriately blamed; to be praiseworthy is to be appropriately praised. Thus, there are norms that tell us when blame is appropriate and there are norms that tell us when praise is appropriate. Undermining factors, we might suspect on this view, show why blame or praise under the circumstances would violate the given norm governing the case. We might hope, then, that the norms governing the appropriateness of blame and the norms governing the appropriateness of praise share the same source. If that were the case, then the Strawsonian would have an answer to the Symmetry Challenge. These factors undermine blameworthiness and praiseworthiness in the same way because they show the response would violate the norms the attitudes constitutive of these notions share in the same way. Just as in the Simple Solution above, we would have identified the common component between blameworthiness and praiseworthiness.

2.3.1. The 'Fittingness of Blame' and Quality of Will

RECALL THAT THE STRAWSONIAN Approach explains blameworthiness and praiseworthiness in terms of the appropriateness of the responses constitutive of blame and praise. Thus, to be blameworthy is to be appropriately blamed; to be praiseworthy is to be appropriately praised. I want to consider an understanding of 'appropriate' that has gained substantial support in the philosophy of emotion, one that might prove helpful to the Strawsonian here.

In their paper "The Moralistic Fallacy: On the 'Appropriateness' of Emotions", Justin D'Arms and Daniel Jacobson explicate a non-moral sense of 'appropriate'. For them, "to call an emotion appropriate is to say that the emotion is *fitting*: it accurately presents its object as having certain evaluative features."[21] Different emotions will present their objects as having different evaluative features. For example, fear presents its object as being dangerous (or something like this). Whether or not fear is appropriate, then, depends on whether or not the object really is dangerous. Pete Graham takes up a similar line with respect to blame. On his view, "[a]ny particular instance of blame...is appropriate just in case the object of that blame…has the features that the blame emotion imputes to it."[22] Just as fear is appropriate just in case it correctly presents its object as being dangerous, blame is appropriate just in case it correctly presents its object as having certain features distinctive of blame responses. The relation between these responses (e.g., blame) and the properties of their objects, then, is analogous to the relation between belief and the world.[23]

Perhaps such an understanding of 'appropriate' can figure into an account of the undermining factors. These factors, remember, render our responses inappropriate. One hypothesis might be that the undermining factors show that the object does not actually possess the features presented by the response. In order to determine the viability of such a hypothesis, then, we need to consider what features blame presents its object as having.

We have a suggestion from Strawson and Wallace. They suggest that our blame-responses are primarily directed at the quality of will the agent's behavior manifests.[24] This suggestion is illuminating. Perhaps blame presents its object as having manifested

[21] D'Arms and Jacobson [2000a], p.65. The italics are theirs.

[22] Graham [2005], p.11. For Graham, the responses constitutive of blame are the emotions of resentment, indignation, and guilt.

[23] This analogy is made explicitly by D'Arms and Jacobson [2000a], p.68. Patricia Greenspan explicitly offers an alternative sense of appropriateness, where 'appropriate' means roughly 'fitting the reasons for the judgment'. This is a radical summary of her view; details can be found in Greenspan [1988].

[24] Wallace [1994], p.128. Wallace is exclusively concerned with blame, but Strawson included "good will" in his characterization.

an ill quality of will. And perhaps praise presents its object as having manifested a quality of good will. The account on offer here suggests that praising and blaming are responses to the quality of will agents manifest towards us. Indeed, this idea is at the heart of Strawson's original proposal.[25] The account on offer here, then, is that the undermining factors show that the agent's behavior didn't actually manifest a quality of ill will or good will. When Fred shoots Barney by accident, it doesn't reflect any ill will towards Barney on Fred's part. And when Bruce accidentally reels in the drowning child we can't take this action to reflect Bruce's good will for the child.

We seem to have a promising Strawsonian answer to the Symmetry Challenge. The same factors undermine blameworthiness and praiseworthiness in the same way by showing that the quality of will blame and praise present their objects as having was not really there. Therefore, blame and praise are inappropriate, and blameworthiness and praiseworthiness are undermined. But there is a problem with this answer. It fails to capture the full symmetrical operation of the undermining factors. This is because the undermining factors affect even cases of non-moral praiseworthiness and blameworthiness. In these cases, no quality of will is necessary in order for the agent to be praiseworthy or blameworthy. To see this point, however, requires showing there to be non-moral instances of praiseworthiness and blameworthiness, and that the undermining factors symmetrically operate even in these cases. I turn now to that task.

[25] Compare Strawson [1962], p.63: "…it matters to us…whether the actions of other people…reflect attitudes towards us of goodwill, affection, and esteem, on the one hand or contempt, indifference, or malevolence on the other."

2.3.2. The Symmetry Challenge Deepens

THUS FAR, WE HAVE BEEN chiefly concerned with the moral appraisal of agents. So we talked of moral blameworthiness and moral praiseworthiness. We saw that the same kinds of factors that undermine moral blameworthiness do so for moral praiseworthiness as well. Thus, I claimed, we should want an explanation of this symmetry that also shows why these factors undermine our responses *in the same way*. This is the Symmetry Challenge in its first instance. I now want to suggest that the challenge can be expanded in a way that further undercuts the ability for Strawsonian accounts to meet it.

First, there is at least one kind of non-moral appraisal that is nonetheless similarly susceptible to the undermining factors. The fact that the practices involved in this mode of appraisal are themselves non-moral, and therefore different than those discussed above, deepens the set of related phenomena any satisfactory account of the undermining factors must explain. The worry for Strawsonian accounts is that as the set of relevant practices grows larger, the prospects for distilling a common feature to explain the symmetry *out of those practices* grow smaller.

Second, it turns out there are actually *many different* kinds of non-moral appraisal, and that they *all* are susceptible to the undermining factors. The fact that a large set of diverse praising and blaming practices are all similarly undermined by the same factors at the very least suggests that the burden placed on the Strawsonian Approach to be able to account for such widespread symmetry out of those practices is severe indeed. And, I think, this evidence places Strawsonian accounts in the position of bearing such a burden, since the Simple Solution reached in Section 2.3 can be marshaled to again give a simple solution to this further evidence.

2.3.3. Artistic Praiseworthiness and Blameworthiness

WE HAVE ALREADY SEEN instances in which moral praise is warranted. But there are also cases in which non-moral praise seems appropriate. As Wallace himself notes, we sometimes praise an artist's "striking and successful work of art," and in so doing, "our praise and admiration reflect a kind of credit on its creator."[26] Like the moral cases, these cases involve taking a stance toward the agent, one that opens him "to direct assessment in virtue of the qualities reflected in the work."[27] We might call this particular case an example of 'artistic' praiseworthiness. If Huckleberry paints a majestic landscape, it seems as if we can praise him for his artwork. Of course, we don't morally praise him. Creating a beautiful piece of art does not make one a morally good person. But it does, it would seem, make one an artistically good person, at least as exemplified by that work. So, we can say that Huckleberry is artistically praiseworthy in this case.

Of particular interest, however, are the reasons for which we are apt to *withhold* our non-moral praise.[28] Suppose we discover that Huckleberry painted his landscape while sleepwalking. I take it we'd be far less likely to praise him for it. In particular, while the painting itself may be no less aesthetically pleasing, it would not seem to reflect on him as an artist in the same way. The painting seems the result of entirely involuntary conduct. And it is this involuntariness which seems to undermine Huckleberry's *artistic* praiseworthiness. Indeed, the painting's quality no longer reflects on him as an artist.

[26] Wallace [1994], pp.53-54.

[27] Wallace [1994], p.54.

[28] Completeness would dictate that I show the effects in each case of each type of undermining factor, but in the interests of space, I limit discussion in each case to a single type of undermining factor. I leave it as an exercise for the reader to construct the other examples.

We could run a similar example for artistic blameworthiness. Wallace notes that we can "condemn the pianist's latest performance…in a way that reflects discredit on the pianist, without blaming the pianist morally."[29] But as above, it would be inappropriate to blame the pianist if his poor performance was accidental. Perhaps the piano was knocked out of tune, and so while his performance was poor, it doesn't reflect on him poorly as an artist.

Wallace admits that we engage in artistic appraisal. But he claims that "…this kind of direct appraisal does not seem especially moral in its quality."[30] From this he concludes that the artist is not morally responsible for his work, though he is "deeply" responsible, in some significant way. Now, surely Wallace is correct to note that our appraisal of the artist in virtue of his work is non-moral. But we should hesitate to draw too strong a conclusion from this fact. Instead, the important observation to make is that even this sort of non-moral appraisal is affected by the undermining factors. The fact that the action was involuntary or an accident mitigated our responses to the artist's work in each case. And the undermining factors seemed to affect our appraisal in the same way in each case. The evidence suggests, therefore, that an adequate account of the undermining factors will not only have to explain why undermining factors make our practices of moral praising and blaming inappropriate, but also why it renders our practices of *artistic* praising and blaming inappropriate as well.

It follows that the Symmetry Challenge can be extended to instances of artistic appraisal. And this extension ought to worry proponents of the Strawsonian Approach. For the non-moral character of the practices associated with artistic praise and blame

[29] Wallace [1994], p.54.
[30] Wallace [1994], p.53.

introduces an increased diversity among the range of practices from which such accounts seek a common feature with which to answer the challenge.

2.3.4. Other Non-moral Modes of Appraisal and Undermining Factors

THE WORRY FOR STRAWSONIAN accounts, however, is larger than the above section suggests. There the worry was generated by observing that there is a non-moral mode of appraisal that the undermining factors nonetheless upset. Thus, the Symmetry Challenge demands that a unified account be given of the undermining factors' effect on both moral and artistic appraisal.[31] Given the difference in the moral character of the two modes of appraisal, Strawsonian accounts would appear to have difficulty in locating a common feature to explain the symmetry out of those practices. But the problem for Strawsonians is actually much larger. It isn't the case that there is only *one* mode of non-moral appraisal. There are at least several non-moral modes.[32]

In addition to artistic praiseworthiness, there also seem to be cases of 'scholastic' praiseworthiness. If Augie gets an 'A' on his math test, then it seems appropriate to praise him for this. Praising him reflects the fact that he's done something good and we are holding him responsible for it. His test grade reflects well on him as a student. Again, it doesn't make him a morally good person, but nonetheless, he seems, for lack of

[31] It is worth noting that my argument in Section 3.3 depends only on there being *at least one* non-moral mode of appraisal. So, even if it turns out that there was good reason to reject artistic appraisal as a particular mode, the points in 3.3 could be restated using a different mode from this section. I used artistic appraisal both because it seems intuitively plausible that such a mode exists, and because Wallace himself points to it as well.

[32] I do not mean the discussion below to provide an exhaustive list of the non-moral modes of appraisal, but merely select two additional prime examples of such modes. Additionally, as in Section 3.3, I limit discussion of the undermining factors to one per mode of appraisal, in the interests of brevity.

a better term, ‘scholastically’ good on the basis of his test grade. So, Augie is academically praiseworthy for his ‘A’.

Additionally, we sometimes appropriately praise athletes for at least some of their accomplishments. If McGraw eagles the 14th hole at the Masters, we can rightly praise his performance. Once again, this need not involve moral praise of any kind. Such athletic performances don’t seem to reflect on the character of the actors in the way that moral actions do, and yet they do invite us to take certain stances toward their performers. Something like this, I take it, serves as part of the rationale behind Most Valuable Player awards. We acknowledge McGraw’s effort and the fact that he performed ‘athletically’ superbly. So, I take it, McGraw is athletically praiseworthy for his eagle shot.

Now suppose that Augie merely circled answers on the test without thinking. While we might marvel at his luck, praise under the circumstances would be inappropriate. It wasn’t as if he meant to get an ‘A’.[33] The accidental nature of his ‘A’ seems to undermine his *scholastic* praiseworthiness,[34] and the grade no longer reflects on Augie as a student. And suppose that McGraw, for his eagle shot, mistakenly used a 7-iron. He meant to pull out his 4-iron (having misjudged the wind, say), but he grabbed the wrong club. In this instance, praising him seems inappropriate. While the result is positive, McGraw only scored an eagle by mistake. And the fact that it was due to a

[33] This is true, I maintain, since he had no reason to think that any of his answers were likely to be correct. He may have ‘hoped’ to get an ‘A’, but this was a hope precisely because he no doubt believed it probable that he would score poorly.

[34] One might initially object to characterizing his ‘A’ as ‘accidental’. But it surely isn’t the case that he meant to get an ‘A’. He couldn’t have, as he was purely guessing. And the result, his grade, seems unexpected in the way needed to classify it as accidental.

mistake undermines his *athletic* praiseworthiness. The excellent shot no longer reflects on him as an athlete.[35]

Praiseworthiness is not limited to its moral instances. We have seen cases that suggest praiseworthiness along alternative scales of value governed by alternative norms. Thus, Huckleberry is artistically praiseworthy in the sense that his painting reflects upon him as an artist (according to aesthetic norms, say). Augie is academically praiseworthy in the sense that his test grade reflects upon him as a student (according to scholastic norms). And McGraw is athletically praiseworthy in the sense that his play reflects upon him as an athlete (according to athletic norms). We can no doubt see the similarity between these evaluations and those involved in cases of moral praising. If Bruce had saved the drowning child on purpose, then the moral praise we accorded would reflect on him as a moral person. We would employ a moral standard of some sort to evaluate his conduct and, finding it exemplary, praise him accordingly. Similarly, in the non-moral cases, we measure the conduct against some other, non-moral standard; for instance, an artistic, scholastic, or athletic standard. But in each case we are expressing the exemplary nature of that conduct with respect to the given standard. While these happen to be non-moral cases, the process of according praise seems strikingly similar to that of the moral domain. Even more striking, however, is that the undermining factors show such praise to be inappropriate even in these non-moral cases.

We can also include analogous examples of non-moral blame. We might evaluate a student's oral presentation in a way that highlights a lack of scholarly achievement, without opening the student up to moral evaluation. This would be a case of scholastic blame. And we might say that Mr. Tennis-Player's serve was unimpressive in the Men's

[35] Or, if one prefers, as a "golf player."

Final in a way that reflects poorly on Mr. T-P as an athlete without implying anything of his moral character.[36] This would constitute athletic blame.

Just as in the cases of scholastic and athletic praise, the undermining factors operate here too. If the student was supposed to write a report on Mother Teresa, but she wrote it on her mother, Teresa, then her poor report is the result of a mistake. It seems that blaming her as a student would be inappropriate.[37] And if Mr. T-P's poor serve is the result of being drugged by his opponent, then it doesn't reflect poorly on him as an athlete. If we suppose his serve is involuntary, then this undermining factor renders even athletic blame inappropriate in this case. It would seem then that blameworthiness, too, is not limited to its moral instances, and yet the undermining factors have a symmetrical effect here as well.

2.3.5. The Expanded Symmetry Challenge and the Simple Solution for Non-Moral Cases

THESE EXAMPLES SEEM sufficient for drawing out the point. The set of undermining factors on these instances of non-moral appraisal is the same as the set of undermining factors on moral praiseworthiness and moral blameworthiness. The undermining factors don't seem to care whether the appraisal of the action is moral or non-moral in nature; they undermine them all in the same way. It is the fact that something is an accident that

[36] I suspect intuitions here may differ. But where they do, I think the objection will be that blaming in this way *is* moral in nature. This claim, I suspect, is only justifiable if there is a moral obligation to, say, do well in school, or try one's hardest in sports. I am extremely skeptical such moral obligations exist. But even if they do, and thus, academic or athletic blame is moral in nature, then these cases belong in the preceding discussion of moral blameworthiness. Such a reorganization does nothing to hinder my argument here.

[37] This claim comes with certain caveats. For instance, if the student simply didn't read the assignment then we might still blame her for her mistake. I still think she wouldn't be blameworthy for the report itself, but that requires separate argument. And in any case, it seems unlikely one could make the sort of mistake she did *without* reading the assignment.

undermines praiseworthiness and blameworthiness, whether that appraisal is moral or not. Fred's eye-poking, Bruce's child-saving, Augie's 'A' test score, and the pianist's performance, are all instances of appraisal that are undermined by the presence of an accident. Jan's coat-stealing, Diana's medication-giving, McGraw's eagle shot, and the student's report, are all instances of appraisal that are undermined by the presence of a mistake. And, Barbie's hitting, Clark's Heimlich, Huckleberry's somnambulistic painting, and Mr. T-P's serve, are all instances of appraisal that are undermined by the presence of involuntariness. It would appear the undermining factors work regardless of the evaluative norms being applied in each case.

The Symmetry Challenge stated that an account must explain why undermining factors mitigate both blameworthiness and praiseworthiness, and do so in the same way. It was initially assumed that the blameworthiness and praiseworthiness in question were moral in nature. But now it seems as if we can expand the challenge to include non-moral cases as well. The undermining factors render blame or praise inappropriate whether or not that appraisal has any moral content. Thus, we can effectively widen the challenge.

The Expanded Symmetry Challenge: An account of undermining factors must explain why those factors mitigate blameworthiness and praiseworthiness in both moral *and non-moral* cases, and do so in the same way.

Now we can see just where the proposed Strawsonian response fails. On the 'fittingness' view, blame and praise are appropriate when it is true that their objects possess the features these responses represent them as having. The feature blame presents is that the object manifested an ill will. Praise presents its object as having manifested a good will. But this view fails to adequately cover the non-moral cases. If McGraw doesn't make a mistake, and still hits a spectacular shot, our justified praise would not represent his action as manifesting a good will. In fact, he may be smug about the shot, intend it to show how much better he is than everyone else, and all manner of other morally problematic attitudes, and still, our athletic praise would be justified. And given the actual scenario, in which his shot is the result of a mistake, the fact that it's a mistake does not serve to show that he didn't manifest a *good* will. There need never have been a good will to begin with. The non-moral cases are not concerned with qualities of will, thus an account that trades on such qualities won't be able to appeal to that notion for these cases of undermined non-moral blameworthiness and praiseworthiness.

After full discussion of non-moral forms of praiseworthiness and blameworthiness, we can expand the Symmetry Challenge to include the undermining factors' effect in these non-moral cases. Despite the inclusion of a shared component, that blameworthiness and praiseworthiness are undermined when the action fails to manifest a quality of will, the Strawsonian Approach cannot meet the Expanded Symmetry Challenge.

I want to pause very briefly to note that the Expanded Symmetry Challenge, while seemingly a bigger problem for Strawsonian accounts than the original version, poses no special problems for the Simple Solution. In fact, the Simple Solution maintains its appeal in accounting for the symmetrical application of the undermining factors across this broader range of cases. Just as being morally blameworthy was a matter of being responsible for something morally bad, we might think that being artistically blameworthy is being responsible for something artistically bad. An extremely plausible suggestion is that the undermining factors upset in some way this explanatorily prior and independent notion of responsibility, one on which all of these norms of appropriateness depend, and thus, the various forms of appraisal (both moral and non-moral) are inappropriate as a result.

As the difficulties mount for Strawsonian accounts, a practice-*independent* approach to responsibility is left unfazed. Part of the reason for this success is that such a notion of responsibility is insulated in some sense from our practices. But since it is largely the diversity of such practices that poses a special problem for the Strawsonian Approach, perhaps an alternative Strawsonian account can be worked out that avoids the problems enumerated above. I turn now to exploring a possible revised view that is particularly sensitive to these challenges (Section 4.1) and showing why it, too, nonetheless ought to be rejected (Section 4.2).

2.4. An Alternative Strawsonian Account

THE SYMMETRY CHALLENGE insists that any account of the undermining factors must show why the same factors undermine praiseworthiness and blameworthiness in the same way. I have shown why the evidence supporting the challenge suggests that there’s a common component that blameworthiness and praiseworthiness share. I have also presented what I take to be a Simple Solution to the Symmetry Challenge:

blameworthiness and praiseworthiness share a notion of responsibility explanatorily prior to and significantly independent of our practices of praising and blaming. Undermining factors show that the agent wasn't responsible for the thing done, and that explains why blameworthiness or praiseworthiness is undermined.[38] Finally, and perhaps most importantly, I've shown why this practice-independent account can explain symmetrical undermining across *all* types of blameworthiness and praiseworthiness.[39]

But aside from making the initial case for a practice-independent approach, we have seen the problems created for the Strawsonian Approach by our observations of symmetry. In the following section, I want to present an initially promising alternative Strawsonian account, one specifically tailored to address the Symmetry Challenge. In Section 4.2, I'll show why this account ought to be rejected, and what's left for the Strawsonian to do.

2.4.1. 'Holding Others Responsible'

ONE WAY WE MIGHT amend the standard Strawsonian accounts in light of this discussion is by suggesting that our praising and blaming practices themselves belong to a shared practice. Recall that the Strawsonian Approach explains being responsible in terms of holding responsible. To be responsible, on this approach, is to be appropriately held responsible. Perhaps a Strawsonian account can be generated that meets the Symmetry Challenge by showing that praising and blaming belong to a more general practice of

[38] On this view, recall, blameworthiness and praiseworthiness require that the agent was responsible for the thing done, and that that thing done is bad or good, respective to the appropriate appraisal scale. Compare Smart [1961], p.305. Smart distinguishes between "grading" an outcome or action, and "ascribing" it to an individual. We can still grade performances without the agent's being responsible, and therefore properly evaluated, by it. Here it seems I am in agreement with Smart.

[39] I have examined here what I take to be a comprehensive sample and believe the same approach would work for any undermining factor.

‘holding others responsible’. The account of undermining factors, then, would be that it is *this* more general practice that is undermined in the various disparate cases. If such a response existed, it might meet even the Expanded Symmetry Challenge by appeal to a shared component of blaming and praising.

The alternative approach owes us an understanding of what it is to ‘hold others responsible’. One suggestion could be that it means adopting a stance of ‘standing ready to appraise’ the agent in the ways appropriate under the circumstances. On this alternative account, ‘it being appropriate to stand ready to appraise’ is a criterion that must be met first *before* the practices of praising and blaming get their grip. The undermining factors could then be reasons why adopting this stance towards the agent would be inappropriate, and thus praising and blaming never even get off the ground. Such an account seems initially promising because it might provide the fundamental condition on praising and blaming of all kinds (i.e., even non-moral kinds), like the one suggested by the Simple Solution to the Symmetry Challenge. On this alternative account I am considering, then, one is responsible if it would be appropriate for others to stand ready to appraise (i.e., praise or blame) her.

We now have an alternative Strawsonian account that attempts to meet the Symmetry Challenge. It does so by suggesting that there is a general practice that ties together our other diverse evaluative practices. It is this general practice that is susceptible to the undermining factors, and this general susceptibility helps explain why the undermining factors have such a widespread, yet symmetrical, effect.

2.4.2. Remaining Difficulties for the Alternative Strawsonian Account

I HAVE TWO RELATED reasons for rejecting this alternative account. First, unlike typical Strawsonian positions, there's little intuitive appeal for the general stance of 'standing ready to appraise'. Its main attraction lies in its being an answer to the Symmetry Challenge. Second, there is a worry that the general stance is really just a presumptive belief that the agent is responsible. That is, 'standing ready to appraise P' is best characterized as 'believing P is responsible (while open to the possibility that P is not responsible)'. But obviously such a characterization is thoroughly unhelpful to a Strawsonian account in meeting the Symmetry Challenge, since it assumes a notion of responsibility independent of the general practice. I suggest that given these objections, it's left to the Strawsonian Approach to defend an independent practice of standing ready to appraise. I take these concerns up in turn.

The initial attraction for the Strawsonian Approach is that we know so much about praising and blaming. By mediating our analysis of responsibility through our practices of praising and blaming, it allows us to explain a difficult concept in reference to less tricky notions. Metaphysical worries about the conditions on responsibility are largely irrelevant to our practices of evaluative appraisal. This aspect of the strategy is one of the more important contributions of Strawson's essay. We may be led by the truth of determinism to doubt that we have a robustly free will, such as the ability to do otherwise,[40] but our practices of praising and blaming will not be similarly deflated. Therefore, if one were able to explicate a notion of responsibility in terms of our evaluative practices, we might expect responsibility to withstand the threat of

[40] One need only look at the debate over "alternative possibilities" to see how divisive the purported ability to do otherwise is. For lively discussion of the prospects, see Widerker and McKenna [2003]. For an excellent discussion of the ability to do otherwise, see van Inwagen [1975], and Vihvelin [2000].

determinism as well. There’s a methodological appeal here to using a well-understood idea to explicate more challenging notions, and one legacy of the Strawsonian strategy has been to use the norms that govern praising and blaming as a framework for explaining responsibility. Praising and blaming are ordinary practices that we engage in all the time, and these practices are seemingly governed by familiar norms. Consequently, this approach is able to remove much of the “metaphysical mystery” surrounding discussions of moral responsibility. Its simplification and appeal to ordinary notions is one of its chief attractions.

But there’s no such appeal with regards to the general practice of standing ready to appraise. Such a practice seems substantially artificial; we have no clear and intuitive grasp of what it would be to stand ready to appraise an agent. We know what it is to praise someone, or to blame them, and we can tell stories about what these practices involve without direct reference the notions of praise or blame. Strawson’s characterizations, for example, involve reference to sets of attitudes that give shape to these practices. But there is no immediate referent attitude to give shape to the general stance of standing ready to appraise. Blame and praise are appraisals; they involve taking attitudes toward others. Standing ready to appraise, in contrast, is not an appraisal. It’s an ‘almost-appraisal.’ And pointing to any attitudes that might constitute such a stance proves a challenge in its own right.

We might take this general stance to be “seeing the agent as a proper target for appraisal.” This sort of stance might be one way to understand a major strand of Fischer and Ravizza’s view.[41] But their view gives us an account of responsible agency; it marks

[41] Fischer and Ravizza hold that a responsible agent is one who is the apt target of the reactive attitudes. See Fischer and Ravizza [1998].

what sorts of creatures can be “apt targets” for assessment. The view sketches a presumption that agents who are reasons-responsive in the right way can be responsible. But such a view will be silent on explaining the undermining factors. For we don’t suppose that someone whose responsibility is undermined in a particular instance, by accident, say, is thereby rendered a non-responsible agent, more generally. The difference here conforms well to a Strawsonian distinction between, in Wallace’s terms, “exemptions” and “excuses.”[42] Exemptions are considerations that suggest an individual isn’t responsible for anything she does; she can’t be responsible because of some capacity she lacks or due to some pathology. Exemptions act globally, and defeat a presumption that the individual is responsible generally. Excuses, in Strawson’s usage, act locally to suggest that an agent isn’t responsible for a particular bit of conduct, what I am calling undermining factors. But these leave intact a presumption that the agent is generally responsible, that there is no reason to suppose the agent couldn’t be responsible for most of what she does. It follows then, that if the general stance of “standing ready to appraise” corresponds to the global notion of responsible agency that exemptions affect, then this stance is of no help to the Strawsonian for meeting the Symmetry Challenge. For that global notion is not one the undermining factors affect anyhow, so it couldn’t serve as the basis for an explanation for their symmetrical effect across the diverse range of cases presented. Therefore, it seems the Strawsonian must look elsewhere for a characterization of a general stance that can meet the challenge.

[42] See Wallace [1994].

Indeed, the most natural characterization of such a stance seems to involve believing that the agent was responsible.[43] If I believe you are responsible, then I am ready to blame or praise you, depending on the evaluative status of whatever it is I believe you responsible for. I should contrast here that the proposed belief is not a belief in one's global responsible agency, but rather a belief that the individual is responsible *for* the particular object under consideration. But this is an unsatisfactory account. It suggests that an agent is responsible only if it would be appropriate to believe that agent is responsible. But it is entirely unhelpful to explain *being* responsible in terms of the appropriateness of our *beliefs* about being responsible. This is just what it means for beliefs to be correct; that they correspond to facts about the world.

More importantly, for our purposes, such an account cannot make any headway toward answering the Symmetry Challenge. The undermining factors show that the response in question is inappropriate; the agent isn't praiseworthy or blameworthy. In order to accommodate this effect within the current alternative account, we must say that an undermining factor shows that 'believing P to be responsible' would be inappropriate. Now, we can ask, how might an undermining factor show such a belief to be inappropriate? One obvious way it might do so is by showing that P isn't really responsible. But this sort of answer isn't available to the Strawsonian Approach, for the approach is committed to the claim that the way to explain responsibility is in terms of the norms that govern our practices of holding responsible. If the general stance of 'holding responsible' is just 'believing to be responsible', then Strawsonian accounts are in trouble again. For one of the most important norms that governs the appropriateness of

[43] Watson says that, "[t]o regard people as responsible agents is to be ready to treat them in certain ways" (Watson [1987], p.256). While Watson explicitly claims this involves more than just a belief, he is unable (in my view) to sufficiently clarify what such a regard amounts to.

beliefs is that beliefs should track truth. But the truth cannot be determined by our beliefs; we need independent conditions to establish a fact of the matter. Thus, in order to show that ‘believing P is responsible’ is appropriate (i.e., correct), we would need to know whether P is in fact responsible. The undermining factors could show that belief is inappropriate only by way of showing that P isn’t really responsible. But such a conclusion would only support a practice-independent notion of responsibility, and thus isn’t available to the Strawsonian Approach.

If a general response of ‘standing ready to appraise’ is to be successful as a means of answering the Symmetry Challenge, then the Strawsonian owes us a compelling story of just what that stance is. And it must be something other than ‘believing the agent to be responsible.’ Without a clear grasp of a practice of ‘holding others responsible,’ and without a persuasive account of how the undermining factors would show such a practice to be inappropriate, we ought to reject this final alternative Strawsonian account.

2.5. Objections

I’D NOW LIKE TO CONSIDER three objections to the argument I’ve presented here. The first objection concerns my insistence that blameworthiness and praiseworthiness ought to behave symmetrically. The second objects to my claim that an account of the undermining factors fails if it can’t account for the factors symmetrical operation “in the same way.” And the third objection questions whether artistic praiseworthiness is really a kind of praiseworthiness.

2.5.1. Wolf's Asymmetrical Freedom

SOME HAVE ARGUED THAT blameworthiness and praiseworthiness are essentially asymmetrical. For instance, Susan Wolf claims that the conditions on blameworthiness and those on praiseworthiness are asymmetrical, because while blameworthiness requires alternate possibilities, praiseworthiness does not.[44] She claims that assertions like, "Jim *had* to save the child's life," are not mitigations of praiseworthiness, but "testimonies to it."[45] In contrast, one who is compelled to kill (and therefore had to do it) is excused (in part) by his compulsion. If there are important asymmetries, the objection contends, this weakens the Symmetry Challenge's position as a criterion of explanatory adequacy. Why should we demand that blameworthiness and praiseworthiness be explained symmetrically if we have data that suggests otherwise?

While I'm independently skeptical of Wolf's argument, it is plain that this is not an objection to the argument I've given here, but rather a set of alternative data points. So we might construct a further constraint on theories of responsibility: they must be able to explain Wolf's asymmetrical data. But this leaves untouched the symmetrical data I've presented, and a view that met Wolf's criterion but failed the Symmetry Challenge would still leave unexplained significant and substantial data. Moreover, I suspect that Wolf's data doesn't show what she purports it does. This is because I don't think compulsion, as we understand it, really undermines responsibility. But I take up this argument in Chapter 3, section 3.3, where I also discuss coercion.[46]

[44] See Wolf [1980].
[45] Wolf [1980], p.156.
[46] For a different criticism of Wolf's asymmetry argument, see Fischer and Ravizza [1992], pp. 375-380. There they argue that there are certain cases in which an agent is blameworthy even though he couldn't have done otherwise. Thus, in their view, neither blameworthiness nor praiseworthiness requires the ability to do otherwise.

2.5.2. How Much Symmetry is Required?

A DIFFERENT OBJECTION concerns the symmetrical operation of the undermining factors on blameworthiness and praiseworthiness. One might accept such symmetry and still deny that this establishes the Symmetry Challenge. For the challenge claims that accounts of responsibility must explain how the factors undermine blameworthiness and praiseworthiness *in the same way*, and one might question why we need symmetry of explanation to account for symmetry in operation.[47] The objection, therefore, seeks to defend a practice-based answer from my argument by showing that it can account for my symmetrical cases. If the Symmetry Challenge isn't entitled to its requirement of explanatory symmetry, then a practice-based account succeeds so long as it gives the right verdicts in each case. And if we don't require that the explanations be the same, then it would seem that the prospects for such a practice-based success are much better than my argument suggests. So, for instance, perhaps accidents undermine blameworthiness by showing blame would be unfair,[48] and they undermine praiseworthiness by showing that praise would be socially disadvantageous. These are just hypothetical suggestions; they are meant simply to indicate how such a strategy might go about accounting for the symmetry.

In reply, I think it's important to remind ourselves of how vast the symmetrical data I've presented is. It would seem that the undermining factors operate symmetrically on moral blameworthiness and praiseworthiness, as well as on multiple kinds of non-moral blameworthiness and praiseworthiness, including artistic, athletic, and scholastic

[47] My thanks to Manuel Vargas and an anonymous referee for *Ethics* for this criticism.
[48] For such an account, see Wallace [1994].

kinds. A Strawsonian account may be able to explain such symmetry by relying on different explanations in each case, or each handful of cases, perhaps. But then we might reasonably ask what ties these explanations together? The larger the set of practices that are symmetrically undermined by the same factors, the more improbable it seems that there isn't some shared feature of the explanations accounting for such symmetry. Even were such a Strawsonian view able to get the right answers across all these cases, then, it would seem to leave a glaring lacuna in its account, for we would still want some answer as to why these explanations fit together. Without such an answer, it would appear that the Strawsonian account asks us to imagine that the symmetry of operation is merely a coincidence, without any underlying unifying explanation. And this lack would put pressure on us again to reconsider the Simple Solution (or some close cousin of it).

2.5.3. Artistic Praiseworthiness as a Kind of Praiseworthiness

A FINAL OBJECTION SEEKS to distance the non-moral kinds of evaluation from the moral ones. So, one might argue that artistic praiseworthiness isn't really a kind of praiseworthiness; it isn't related to moral praiseworthiness in any interesting way.[49]

But this objection misses its mark. Nothing beyond the evidence presented so far is required to get my argument started. Artistic praiseworthiness, for example, shares at least one interesting feature with moral praiseworthiness: both are undermined by the same considerations. It is, of course, an open question what explains this shared feature. I have presented what I take to be a Simple Solution worth pursuing. Nevertheless, it

[49] This is related to a criticism made by an anonymous referee for *Ethics*, who claimed that it wasn't obvious that non-moral instances shared anything interesting with their moral counterparts besides a notion of "causal" responsibility.

would seem that the simple fact that each is affected in the same way by the same considerations is enough to establish that they are related in some interesting way.

2.6. Conclusion

ANY ACCOUNT OF THE undermining factors must explain why the same factors undermine both blameworthiness and praiseworthiness in the same way. Strawsonian accounts seem unable to meet this challenge. In contrast, an account that stresses a practice-independent notion of responsibility seems able to easily and plausibly meet the challenge. Being responsible is a necessary condition on being blameworthy or praiseworthy, but we should not explain responsibility *in terms* of these two evaluative concepts. Rather, a better prospect would seem to lie with pursuing an explanatorily prior and independent notion of responsibility, along the lines of the Simple Solution to the challenge. We are blameworthy when we are responsible for something bad. We are praiseworthy when we are responsible for something good. To my mind, this is an intuitive proposal to develop, and it can easily accommodate the Symmetry Challenge. I turn now to developing such a proposal, pursuing a thoroughly compatibilist account of an independent notion of responsibility. In the following chapter, I begin by examining a more complete list of undermining factors, and pursue a promising hypothesis that understanding why particular factors undermine responsibility will point us to positive conditions on responsibility.

Chapter 3: The Undermining Factors and the Conditions on Being Responsible

3.1. Introduction

IN THE LAST CHAPTER, I presented a challenge to Strawsonian accounts of responsibility. These accounts are unified by explaining what it is to be responsible in terms of the appropriateness of holding responsible; that is, our practices of praising and blaming. I argued that this is the wrong order of explanation. Instead, I suggested, we need an independent notion of responsibility in order to explain the symmetrical undermining of blameworthiness and praiseworthiness by the same factors across a broad and diverse range of cases. The purpose of this chapter is to fill in what that notion of responsibility looks like, and to begin to show what the conditions necessary for being responsible are.

On my view, one is blameworthy just in case one is responsible for something bad. One is praiseworthy just in case one is responsible for something good. In each case, we have a responsibility component, together with an evaluative component (good or bad). We can account for the various types of blameworthiness and praiseworthiness previously discussed by simply adjusting the norms involved in the evaluative component. So, moral blameworthiness is governed by moral norms, but those norms tell us whether or not the outcome was morally bad. They lie outside of considerations about whether the agent is responsible for the outcome. Similarly, artistic praiseworthiness is governed by artistic norms. But they, again, simply tell us whether or not the outcome is artistically good. The responsibility component in both cases is independently assessed.

This Simple Solution to the Symmetry Challenge is able (intuitively, I think) to capture the shared component between the various cases of symmetrical undermining canvassed in the last chapter. Moreover, it is able to help explain how the symmetry is sustained across the wide range of diverse cases already discussed, because that diversity occurs with respect to the evaluative component only. Thus, the same structure is present in every case; only the set of norms determining the appropriate evaluative status of the outcome in question changes.[50]

An account of this responsibility component should tell us just when an agent is responsible. To do this, I will again look at the undermining factors. If the undermining factors symmetrically diminish blameworthiness and praiseworthiness across a diverse range of cases, and do so in the same way, and if they do so by affecting this independent notion of responsibility (as the Simple Solution suggests), then a plausible hypothesis is that each undermining factor suggests the lack of a condition necessary for responsibility. This hypothesis is initially further supported by the observation that many undermining factors seem to present negative features. They show the outcome was *in*advertent, or *in*voluntary, or *un*intentional, or by *mis*take. It seems promising to suppose that perhaps if we remove the negations, we'll be left with conditions an agent must satisfy to be responsible. This is just an instantiation of a general strategy of explanation. If you want

[50] This is to say that moral norms wouldn't be correct for artistic evaluation, and vice versa.

to figure out what makes a rhombus a rhombus, a promising strategy is to look at shapes that are almost rhombuses. Contrasting rhombuses with these figures will tell you something important about what makes a rhombus such. The strategy here is to look at cases of undermined responsibility to tell us something important about what makes an individual responsible for what she does. In short, the idea is that by investigating the undermining factors, we will thereby be able to distill from their negative features the requirements on being responsible. The natural thought is that these negative features help demarcate the line at which responsibility is undermined; so, if no negative feature is present, the agent is responsible.[51]

My strategy is as follows. The first step is to determine the list of undermining factors, and to characterize the negative features they exhibit. The aim is to achieve a classification of the undermining factors grouping specific factors together into classes by virtue of the negative features they highlight. I begin this characterization in Section 2, with what I take to be the most uncontroversial considerations, with a special emphasis on the ways mistakes operate. In Section 3, I argue against the idea that instances of coercion undermine responsibility. So, while they may initially seem to count as cases of undermining factors, I conclude we ought to exclude coercion from our set. Section 4 sums up what we learned from our characterizations in the previous three sections, and outlines what I take to be the set of necessary conditions on being responsible suggested by the undermining factors. Section 5 concludes by highlighting the connections

[51] This is essentially the Second Prong of Strawson's strategy. Following a suggestion by Gary Watson, Michael McKenna interprets P.F. Strawson in "Freedom and Resentment" as also engaging in this strategy of distilling the conditions on responsibility from the undermining factors. For Strawson's original discussion, see Strawson [1962]. For McKenna's discussion, see McKenna [1998]. The difference, as I see it, between Strawson's approach and mine is that I explicitly pursue positive conditions distilled from categories of negative features, whereas Strawson assumed responsibility if no undermining factor was present.

between the conditions adduced here and the Simple Solution to the Symmetry Challenge from Chapter 2.

3.2. Paradigm Cases

WE NEED ONLY CONSIDER our everyday experiences to reveal a large variety of considerations offered to excuse blame or deny praise. The simplest expressions of undermining factors take the form of ‘I didn’t mean to’ or ‘She didn’t know’ or ‘He couldn’t help it’ or ‘It wasn’t my fault.’ Virtually everyone has made at least one of these claims (I admit to all four…) in an effort to mitigate blame from others. Similarly, we often point to the same considerations when arguing that some praise is undeserved. We tend to say things like ‘But he didn’t *do* anything’ or ‘She was just lucky’ or ‘It was just a happy coincidence.’

These phrases all point to considerations that are meant to show that the agent isn’t blameworthy or praiseworthy because they aren’t responsible for the outcome for which the agent is being blamed or praised. They are the language of the undermining factors with which we are most familiar. It should be noted that both the enormous complexity of even ordinary events and the flexibility of ordinary language make it difficult at times to pinpoint what particular undermining factor was present or is being pointed to. It is often the case that several factors are present at once, or sometimes it proves difficult to cite just what the factor was, even when we’re certain responsibility in such a case is undermined. In the normal course of things it often does not matter whether we can cite the particular undermining factor. But there are paradigm cases, and it is to these that I shall devote most of my attention. In addition, I think it often the case that we can pick out more precisely the undermining factor at play, even in complicated cases, and we should endeavor to do so even if it proves particularly challenging at times

At any rate, I think it clear that when one consults the range of considerations we are likely to offer in our defense against various accusations, or as reasons that praise is inappropriate, a natural classification begins to emerge. There is, of course, always a worry that such a list won't be exhaustive. But the goal at this point is to chart a categorization of the paradigm instances of factors that undermine responsibility. I'm willing to sacrifice some exhaustiveness for greater progress towards that goal. In any case, I do think the undermining factors can be grouped into a few main categories. And while I don't suppose to discuss *every* possible individual undermining factor, I will discuss these main categories of factors (and I do take the *categories* to be exhaustive) using illustrative example cases.

3.2.1. The Involuntary

ONE SET OF CASES that rather obviously undermine responsibility involves instances of involuntariness.[52] If my wife rolls over in the night, knocking me out of bed, she is not responsible for my rude awakening. Recall the example of Huckleberry, who painted a majestic landscape while sleepwalking. He is not responsible for the gorgeous painting.[53] These are outcomes produced while the agent in question was unconscious. Similarly,

[52] Throughout, my contrast case (to the moral) will be cases of undermined responsibility with respect to artistic outcomes. My aim is to keep the discussion simple and centered. Nevertheless, I think we could construct as many parallel examples as we wanted with other normative outcomes (e.g., athletic, scholastic, etc.)

[53] Recall that here, being responsible for the painting means that the painting would reflect on Huckleberry as an artist. We still may marvel at his strange "ability," but we wouldn't consider the painting as an "accomplishment" of his as an artist.

suppose Barbie suffers a seizure and hits Ken. Or suppose that Carter suffers a seizure, and in so doing draws a remarkable portrait. These are instances of spasms, and in neither case, I think, is the agent responsible for the outcome. We could also construct related cases where the outcome is produced via a reflex of the agent, such as knocking over a vase while swatting away swarming bees. There are also, it seems, related cases where one fails to do something as the result of incapacitation (but not unconsciousness). Examples here would be where one fails to meet an associate for lunch because one has been bound and gagged, or where a musician's performance suffers from an intermittent, but persisting paralysis in his hands.

In all these cases, it seems as if the agent's responsibility is similarly undermined. And it seems as though what ties these cases together is that the outcome is produced involuntarily. In each case (save incapacitation),[54] the agent performs bodily movements that produce an outcome, but those bodily movements themselves are involuntarily produced. And we can certainly contrast these cases with the ordinary actions of our colleagues and friends, when they hit others or fail to keep a lunch date.[55]

3.2.2. The Unintentional

ANOTHER SET OF CASES seem to highlight that the outcome was unintentional in one way or another. For example, recall the example of Fred and Barney, in which Fred opened a bottle of champagne, and the cork popped, bounced off a wall and hit Barney in the eye.

[54] The case involving incapacitation is excluded only because it involves the lack of bodily movement, not because it isn't involuntary produced.

[55] In this way, my project is compatible with the Strawsonian project of naturalizing responsibility. Where I differ most strongly with the Strawsonian tradition is in my explanation of the notion of responsibility – what it is to be responsible. It is this disagreement that was the focus of Chapter 1.

This is an incident in which the outcome is an accident. Compare this example to the case of Augie, who merely circled answers on the test without thinking, and yet scored a perfect grade. Here, again, the result is an accident. And in neither case, I think, is the agent responsible for the accidental outcome. Here, it seems that the fact that an outcome is an accident serves to undermine responsibility.

A related set of examples concerns inadvertence. Suppose a bunch of friends have gathered in the living room and are watching a movie. When Lenny gets up from the couch to get a soda, he's distracted by the dialogue and so he steps on a friend's hand. It seems as though Lenny's responsibility is undermined because he only steps on the hand inadvertently. Similarly, imagine that while Marta is putting the finishing touches on her painting, her friend enters the room and says "hi." When Marta turns to wave back, her brush sweeps across the canvass, creating a remarkable effect on the painting. Here, again, it seems as though Marta's responsibility for the painting is undermined, because she only makes it remarkable inadvertently.[56]

Inadvertent outcomes are side-effects of intentional action, where the agent fails to foresee the possibility of the side-effect, but would foresee it if the agent considers his situation more carefully. If Lenny is mindful about looking at the floor, taking care to where each step will land, etc., it is likely that he will avoid treading on any hands. Accidents are also side-effects of intentional actions, but whereas the inadvertent could have been foreseen with additional attention, accidents are unexpected or unforeseen, even had reasonable consideration of the possible consequences been given. While it may not have been absolutely impossible for Fred to foresee that the cork could hit

[56] We can suppose that responsibility for the painting concerns the finished painting, especially with respect to those features of it that make it remarkable. Obviously, Marta is likely responsible for all the aspects of the painting up until the inadvertent brushstroke (if anyone is ever responsible for outcomes in the world).

Barney in the eye (Fred may be able to imagine possible scenarios when considering opening the bottle), the resultant outcome is not among the list we might make of the possible outcomes within the realm of likelihood. Accidents, therefore, seem to involve a notion of happenstance, whereas inadvertent outcomes are the result of a lack of attention. This is a rough characterization, but we do seem to readily distinguish accidents as 'unlikely outcomes,' from inadvertent outcomes, which don't exhibit the same degree of unexpectedness. At the very least, we can easily note that Fred's eye-poking is an unexpected consequence of his bottle-opening in a way that Lenny's hand-stepping was not an unforeseen consequence of his getting up. And this seems enough to outline the distinction. We may, of course, call certain outcomes "accidents" that weren't really unlikely. Car accidents, for example, may be the result of dangerous driving that greatly increased the chance of a crash. Despite this usage, it seems that cases like Fred's are closer to "core" examples of accidents, outcomes that are actually unexpected or reasonably unforeseeable given the circumstances under which the agent acted. Such core cases, it seems, are the ones to look to if we are to properly characterize what is central about such types of cases. Core cases would seem to be most illuminating and reflective of what is different in cases of accident than in paradigmatic intentional action.

A further case seems to involve ignorance. Ignorance can also undermine responsibility. Suppose Martin walks to his room and opens the door, hitting his roommate Mark in face. Mark had been admiring his new outfit in Martin's full-length mirror, but Martin didn't know that Mark was there. So Martin didn't know that his opening the door would harm Mark, and this ignorance seems to undermine his responsibility for the injury. Now imagine that Luigi doesn't know that his guitar has been wired such that every time he plays the third string, 15th fret, his amplifier will produce a sympathetic tone. Suppose that his otherwise respectable solo is rendered remarkable by this effect, which happens intermittently throughout the performance (indeed, just those times when Luigi plays the 3rd string, 15th fret). Here, too, responsibility seems undermined (at least with respect to the solo's remarkable aspect), and it seems to be Luigi's ignorance that is doing the undermining work. To make this point clear, remember that while we might grade the solo itself as remarkable, it would seem a mistake to attribute this quality to Luigi as a guitarist.

In all of these instances the outcome is unintentional or unforeseen. While the agent performs some action intentionally (e.g., opening a bottle, marking an answer, walking to the kitchen, turning around, opening a door, playing a solo), the specific outcome in question, the one we're interested in, was not part of the 'intentional structure' of that action. It wasn't an aim or end of the action, it wasn't a necessary step towards an intended goal, nor was it a foreseen side-effect.[57] We are all too familiar with cries of "But I didn't mean to" when faced with a criticized outcome. This general phrase seems to point us to considerations of accident and inadvertence, in order to show that the outcome in question was unforeseen, and that responsibility is therefore

[57] Inadvertent outcomes, we might think, are foreseeable even when they are not actually foreseen by the particular agent. Accidental outcomes, on the other hand, might be distinguished by their 'reasonable unforeseeability'; i.e., that it would be unreasonable to expect an agent similarly placed to foresee the actual outcome of the action. Here I merely want to characterize how we seem to distinguish between these considerations. It seems that claims of inadvertence serve to show the actual agent didn't actually foresee the outcome, whether or not he could (or should) have.

undermined. Similarly, "But I didn't know" references the lack of relevant beliefs to the same effect.[58]

3.2.3. The Mistaken

THERE IS ANOTHER SET of factors that undermine responsibility through reference to an agent's beliefs.[59] Where ignorance picks out the lack of relevant beliefs, mistakes reference the agent's false beliefs. For example, recall the case of Jan taking a coat she believes to be hers as she's leaving a party. As it turns out, the coat she takes belongs to someone else, though it is the same type, size, and color as hers. Here Jan is mistaken about whose coat it is she is taking. She believes it to be hers, when in fact it is not. But because intentions are sensitive to beliefs, and because Jan believes the coat is hers, she doesn't intend to take someone else's coat. She intends to take her own coat, and simply fails to fulfill her intention, due to her false beliefs. It also appears that Jan's responsibility for stealing a coat is undermined due to the fact that she made a mistake.[60] Similarly, suppose Jimmy pulls out the wrong brush. He thinks he's chosen a fan brush, but it's really a #3 pointed brush. Despite his error, when he uses the brush it creates a remarkable effect on the painting. As in Marta's case, even if we grant that Jimmy is responsible for the rest of the painting, it seems his mistake undermines his responsibility for this particular remarkable effect.

[58] We sometimes use "But I didn't know" to cite a mistake. As will become clear in the following category, mistakes nonetheless deserve separate treatment. This is just one example of why we must seek greater precision than our ordinary phrases provide.

[59] In some cases, the false belief may be attributed to another kind of "mistake"; for example, a 'misjudging', 'mishearing', or 'miscalculating'.

[60] One might like to qualify the statement here and say that Jan makes an "honest" mistake. But such qualifications do not suggest a criterion on mistakes (i.e., that only mistakes that count as "honest" serve to undermine responsibility), but rather speak to our confidence that it was an actual mistake that was made.

What distinguishes mistakes from accidents and inadvertence, or even ignorance (where one fails to have the relevant beliefs), is that mistakes do not show the outcome was unintentional, full stop. Part of the reason is that in many cases of mistake, the outcome in question is not a side-effect of the intentional actions of the agent. Unlike Martin's case, where Mark's getting hit in the face is a side-effect of Martin's opening the door, Jan's taking the coat is both the action and outcome in question. As a result, Jan surely takes *the very coat she takes* intentionally. She meant to pick up that particular coat, carry it with her outside, put it on, and drive home. But when she appeals to her mistake, she makes reference to her belief, sustained throughout that series of actions, that the coat in question belonged to her. As such, citing a mistake is an effort to *specify* the intention the agent acted on; mistakes highlight what the agent believed herself to be doing. While Jan intentionally took *that* coat, she didn't intentionally take *someone else's* coat. And we can explain the difference by referencing Jan's mistaken belief that it was her coat, for this fact shows both why she intended to take that very coat and how she did not act from the intention to take someone else's coat.[61] In this sense, a claim of mistake seeks to show that Jan took someone else's coat unintentionally.

As a class, then, the Mistaken is similar to the Unintentional, in that considerations in both categories can show that the outcome in question was unintentional (or unforeseen). But it is important to treat mistakes separately, if only because the only way to show that the outcome was unintentional (or unforeseen) is by reference to the agent's false beliefs. In Jan's case, the only way to show that she takes someone else's coat unintentionally is to reference her false belief that in acting she was

[61] This is just a claim about the opacity of the content of our intentions and beliefs. For example, Lois can intend to meet Clark Kent for lunch and fail to intend to meet Superman for lunch, despite the fact that Clark Kent is Superman. Similarly, it would seem Lois does not intentionally have lunch with Superman.

taking her own coat. As Jan's case illustrates, mistakes are characterized by an agent's false beliefs about facts that translate into false beliefs about the nature of what the agent is doing. Jan doesn't realize she's taking someone else's coat because she believes that the coat is hers.

I've set out what I take to be paradigm instances of the undermining factors (the Involuntary, Unintentional, and Mistaken). From these we have gotten a picture of the ways in which responsibility for outcomes can be undermined. It can be undermined when the agent's action was involuntary. It can be undermined when the outcome was unintentional, either as an unforeseen side-effect or as a result of false beliefs. I take these categories to be exhaustive of the different types of undermining factors. But before turning to distilling the necessary conditions on responsibility from these categories, I need to address a possible concern, namely that I have left important potential undermining factors out. Chief among possible candidates are instances of coercion, which are often taken to undermine responsibility.[62] In the following section I argue that this suggestion is wrong, and that instances of coercion do not undermine responsibility. If I'm right, then coercion does not belong in our list of undermining factors, and we can proceed with isolating the necessary conditions on responsibility.[63]

[62] For just one such example, see Wallace [1994].

[63] As noted earlier, one obvious drawback to my methodology is that we can't be certain we've looked at every possible undermining factor. While we may be barred from absolute certainty, I know of no place where a typology of undermining factors is discussed that includes a consideration that isn't represented by my categories above, or else dismissed as an undermining factor (as I intend to show with respect to coercion). I think we should be highly confident, therefore, that the account we're led to by these categories will not be significantly lacking. For examples of such typologies (or approximations thereof), see Wallace [1994]; Austin [1957]. One possible exception is the excuse "I couldn't have done otherwise." I discuss such a consideration at length in Chapter 4, Section IV.5.

3.3. Coercion

I CAN NOW TURN TO CONSIDERING cases of coercion as potential undermining factors. It can seem natural that situations of coercion undermine responsibility. Suppose Moira holds a gun to Manny's head, demanding that he crack a safe. Supposing that Manny complies, it appears Manny isn't responsible for cracking the safe, and this is due to Moira's coercion of Manny. Or suppose that Malek threatens Micah's family with harm, if Micah doesn't lie about Malek's whereabouts. Here, too, it initially appears as though Micah isn't responsible for lying. In neither case, for example, are we liable to blame the coerced agent. So, we might ask, what is it about coercion that undermines responsibility? One observation to make about both cases is that the agent is presented with an especially powerful incentive for the choice they make. In Manny's case, Moira threatens him with death, a surefire incentive for most of us. In Micah's case, Malek threatens the welfare of his family, another forceful incentive.[64]

Of course, naturally, one often says in such situations, "I had no choice!" But this is meant more as hyperbole than as fact. It isn't that Manny has no choice, only that he has no *good* choice. If he refuses, he gets shot in the head. If he accedes, he helps rob a bank. Neither is a particularly attractive option. Furthermore, making reference to a reduced capacity to choose is meant to indicate that the options from which one could choose are not endorsed by the agent. If Manny were able, he'd refrain from either alternative (death or stealing). Similarly, Micah has a choice. He could refuse to lie, but

[64] Of course, Manny and Micah need not themselves feel the "pull" of these incentives. Manny may be so principled so as to never do what he believes to be wrong, even when threatened with death; and Micah may be radically disconnected and feel nothing for the welfare of his family. That such people may exist, to my mind, does little to discredit the observation that, in general, ordinary folks would feel the requisite pressure presented in the cases of Manny and Micah. In any case, I doubt we'd consider cases without such pressures as instances of coercion.

only at the cost of his family's welfare. Or he can choose to lie. Again, he has a choice, just not a particularly good one. And similarly, his preference would be none of the alternatives on offer.

Cases such as these may suggest to some that the agent does not act voluntarily. But such a claim suffers from vagueness. After all, there is a perfectly respectable sense of 'voluntarily' such that both Manny and Micah act voluntarily. Each makes a choice, based on the available reasons, and carries out an action based on those reasons. We act in this way all the time. I can't have both the pizza and the eggplant parmesan; so I pick one. The difference for Manny and Micah seems to be that the choice in this instance is, in some sense, forced, *and* that neither choice is attractive. One might object that little bars me from choosing both the pizza *and* the eggplant parmesan, and I could at least try to get both. But Manny's case is similar in this respect, since we can well imagine him rejecting the constraints of the forced choice, and instead trying to wrestle the gun away from Moira, or running away. Micah may be in a more difficult spot, but he too, it seems, could have recourse to options "not on offer" as I've set things up. These observations lead me to suspect that in acting under coercion we act far more voluntarily than is typically admitted.

Now for all I've said, coercion *could* still undermine responsibility. Perhaps it really is necessary that the agent must find the options themselves to be appealing or endorse them in some way in order to be responsible for the outcome. But I'm skeptical. And I think there's good reason to doubt that coercion really undermines responsibility. My strategy in defense of this claim is to argue that cases of coercion are structurally similar to cases of *necessity*, and that both sets of cases should be treated similarly. And since I think we should be more confident that necessity does not undermine responsibility than that coercion does, we ought to exclude coercion from our classification of undermining factors.

3.3.1. The Argument for Excluding Coercion

MY ARGUMENT FOR EXCLUDING coercion begins by setting out a common distinction between two types of defenses to ascriptions of blameworthiness: excuses and justifications. These defenses are standard practice in the criminal law, but they are readily adapted for use here.[65] An excuse claims that what was done was in some way bad, but that the agent isn't responsible for it.[66] A justification, however, allows that the agent is responsible for the outcome, but suggests that the outcome itself was (in some sense) good or proper. A justification claims it is proper to do as one did "under the circumstances" or "given his options."

The next step in the argument is the observation that coercion is often thought to excuse agents, while cases of necessity are thought to be justifications.[67] Now, a common example of necessity from legal thought is self-defense.[68] So, consider Nora, who is being attacked by an axe-wielding Nick. The only way she can defend herself is by incapacitating him, which she can accomplish by pulling a lever that drops a very

[65] Indeed, the criminal law is a valuable resource in thinking especially about blameworthiness. Criminal liability is usually taken to be a stringent standard necessary for justifying punishment. Since many believe blameworthiness necessary for justifying blame, it seems that criminal liability shares something in common with (moral) responsibility. Both are taken to provide a grounds for evaluation based on the relevant standards; morality on the one hand, and the criminal code on the other.

[66] See Austin [1957]; Rodin [2002]; Wallace [1994]. 'Excuse' as I'm using it here, then, is a somewhat technical notion, since we sometimes count as "excuses" considerations that don't aim to remove responsibility, but seek only to mitigate others' responses to our conduct, often to ameliorate potential sanctions. I have more to say about this in the next section, and in Ch. 4, Secs. IV.2-IV.3.

[67] For an excellent discussion of the distinction and pointers to further reading, see Rodin [2002].

[68] Model Penal Code, Article 3, §3.04.

heavy weight on him. She pulls the lever, the weight is dropped, and Nick is incapacitated. Here it is natural to think that Nora is responsible for injuring her attacker. Indeed, what she did was appropriate, given the circumstances, and we would expect Nora to pull the lever and think it the proper step to take to save her own life.[69]

But notice that Nora's case is structurally similar to Manny's. Manny had to choose between cracking a safe and getting shot in the head, and neither of these is an attractive option. Moreover, we supposed, Manny would prefer it if he could avoid making the choice altogether. Similarly, Nora must choose between being chopped to death and crushing Nick, neither of which by itself is particularly attractive.[70] And she, too, would likely prefer to avoid the choice entirely. Given that these cases are structurally similar, I think there is pressure to treat them similarly. So, it ought to be the case that coercion and necessity are to be treated either (1) both as excuses, undermining responsibility; or, (2) both as justifications, leaving responsibility intact. Establishing this disjunction is step three.

Step 4 makes the case that option (2) is preferable. Why should we prefer treating coercion and necessity as justifications? My answer is that it is implausible to suppose that instances of necessity undermine responsibility. To see this, we need only consider a few examples. Nora's case does not appear to be one in which she isn't responsible. On the contrary, she seems perfectly responsible – and for doing the proper thing. We can

[69] It is certainly possible for some to react differently to the case, and claim that Nora doesn't act appropriately. But in such cases I suspect one would also concede that Nora is blameworthy for her act. A related, but distinct, point concerns individuals who might think harming even in self-defense is inappropriate and refuse to do it (e.g., Quakers or pacifists). Here, too, we can agree that refusing to harm in self-defense may as well be permissible. Such a conclusion does not modify the claim that a Quaker would be responsible for his refusal.

[70] That is, there are significant reasons against each option. Crushing Nick may be "attractive" to Nora in that she would take delight in it, say, or that it would save her life. But that isn't the sense I'm interested in. "Choiceworthy" reflects the sense of "attractive" at issue here.

also compare another standard case of necessity. There is a wildfire heading straight for the center of town, and Otis notices that it will pass through his neighbor's field on the way.[71] So, Otis burns down his neighbor's field, exhausting the fuel the wildfire would have otherwise consumed, and preventing the fire from ravaging the town. Here, Otis surely is responsible for setting fire and burning down his neighbor's field. He does it deliberately, knowing all the relevant facts. This is a paradigm case of responsibility. Now, of course he acts for good reasons. He only burns down the field so as to save the town from much greater harm, but this reasoning is precisely what justifies his choice to us. Far from undermining his responsibility, the necessity of the situation is his explanation for why he brought about what he's responsible for.

Still, we need not engage in such serious scenarios to see the point. We encounter similar situations all the time. For instance, tonight I could clean out the garage, as I know that its current state is distressing my wife, or I could do something which I would enjoy more. But I know that if I fail to clean the garage, it will continue to cause my wife distress, which would be worse than missing out on my enjoyment.[72] In these situations it seems as though I am responsible for the outcome I choose.[73]

By parity of reasoning, coerced agents, like Manny, choose well, and are responsible for the outcomes they bring about. Manny, too, has an explanation for why he cracked the safe. He wanted to avoid a worse harm: his death. This is his defense against our recrimination, but it doesn't appeal to undermined responsibility. Instead, it

[71] This standard case of necessity is perhaps best known as a "lesser of two evils" case. When one chooses the lesser of two evils, it is thought, one chooses appropriately (provided one cannot avoid the choice).

[72] The distress she suffers need not itself be worse than the loss of enjoyment that I suffer. It could be that causing my wife distress, or allowing her to be distressed when I could easily prevent it, is worse than the loss of enjoyment I suffer by missing the game.

[73] Similarly, Aristotle took coercion and necessity to be of a kind, and that they amounted to justifications provided the agent picks the "choiceworthy" action. For his discussion, see Aristotle [1999], Bk. III, Ch. 1, esp. sec.1-6.

points to those considerations that Manny thinks explains why he didn't refrain from doing what he had good reason not to do. Just as Otis has strong reasons not to burn down his neighbor's field, reasons that are outweighed in his particular case, so too does Manny have good reasons that are simply outweighed in this instance. Otis is justified so long as we take the reasons he gives to be strong enough to make his choice the appropriate one under the circumstances. Manny is justified if the same condition is met.

So, I think we do better understanding both coercion and necessity as instances of justifications, which do *not* undermine the agent's responsibility. It is implausible to think that for all the various cases of necessity that we can construct the agent's responsibility is undermined. This concludes my argument for treating coercion in the same way as necessity, that is, as a justification.

3.3.2. Objection 1: Coercion and Blameworthiness

ONE MIGHT OBJECT to this conclusion in the following way. If I am right about coercion, then this suggests coerced agents are responsible for the outcomes they bring about. But cracking a safe isn't a good thing; it is instead a bad thing. And according to my conditions for blameworthiness, this would imply that Manny is blameworthy. But coerced agents aren't blameworthy. Similarly, if Otis and Nora are responsible, then they are responsible for some bad outcomes, which means that they, too, are blameworthy. But justified agents aren't blameworthy. So my account of coercion (and necessity) must be wrong.

Notice first that this objection does not affect my claim that coercion and necessity should be treated alike; it only states that neither constitutes a justification.[74] I find this in general to be an implausible claim. But what should we say about the claim that Manny is blameworthy for cracking the safe? Or about the claim that Otis is blameworthy for burning the field?

Let's evaluate these claims by examining some tougher cases of coercion.[75] Suppose that when Moira threatens Manny's life, she doesn't want him to crack a safe, but to kill 10 other people. Here it is much harder to say that if Manny chooses to kill them he acts appropriately. But now suppose that all Moira wants is for Manny to kill 2 other people. In this case it may still be difficult to say Manny chooses appropriately if he kills the 2 people, but it seems to be a more likely case of a justification than in the '10 kills' variation. Suppose Manny does kill the 2 in order to save his life. Is he justified? First, I want to dismiss a potential worry about such cases that has already been addressed above. Conceding that Manny is justified here does not mean that killing the two is a good thing. Justifications do not concern what "ought to be the case," but rather what "ought to be done." Given this requirement, Manny is justified so long as killing the two is what he ought to do, *given the circumstances*. Manny of course may not be justified in killing the 2 in order to save his life. The answer to whether or not he is in fact justified lies with our substantive moral theories regarding cases like his. I do not suppose to be able to weigh in on such a question here. I will simply remind the reader that no matter the answer, it will not change Manny's responsibility for the deaths, and that coercion is therefore not an undermining factor.

[74] Additionally, the objection might conclude simply that my explanation of blameworthiness requires revision. In any case, my reply below answers both prongs of the objection.

[75] Everything I say here can be modified to apply to necessity simply by adjusting the examples a bit.

Second, notice that the claim that Manny is responsible is compatible with either the existence or non-existence of moral dilemmas. On at least one plausible interpretation, a moral dilemma is a situation in which no matter how an agent chooses, he chooses wrongly. No matter how we settle on the question of whether or not there are scenarios in which there is nothing an agent can do permissibly, or, where every option is one that the agent ought not do, this will not affect the above formulation of justifications. That is, the answer to dilemmas will not affect whether or not coercion can be a justification. It will only show that in certain cases, Manny chooses poorly, and is blameworthy as a result, though no choice he could have made would have been appropriate. Judgments regarding the plausibility of such a dilemmatic scenario regard the acceptability of dilemmas themselves, not about whether coercion is a justification or not.

Finally, I suspect that claims to the effect that Manny isn't blameworthy for killing the two are driven by two important considerations: (1) that someone else is blameworthy in the situation, namely, Moira; and, (2) that we have decisive reason not to actually blame Manny. On the first point, Moira is clearly blameworthy. She is responsible for threatening Manny, and that's morally bad. Additionally, it seems as though she surely wants the two people dead, and her threatening Manny is an attempt to get him to fulfill that aim. She's using the probability that a convincing threat will motivate Manny to achieve her ends, and those ends are also morally bad. Moreover, it seems that Moira, too, shares some responsibility for the two deaths. After all, she is trying to bring something about in the world, these two deaths, and shapes her actions to influence Manny into helping her bring that about. But this last fact does not absolve

Manny from responsibility. Moira's attempt to influence Manny could fail – if Manny refuses. And then neither could be responsible for the deaths. It takes Manny's agreement (of a sort) to fulfill Moira's aims, and to do so Manny must share that aim (at least to the extent that it saves his life).[76]

On the second point, just because an agent is blameworthy does not settle the question of whether or not we ought to blame him. To deny this distinction is to accept what D'Arms and Jacobsen have called "The Moralistic Fallacy."[77] Something can be fearworthy even though you shouldn't fear it (because doing so will make it more likely that you'll fail some important aim, say). And similarly, just because you ought to fear something, doesn't mean it is fearworthy (perhaps if you genuinely fear something you will be rewarded with immense wealth). I believe blameworthiness to work the same way. So while the fact that an agent is blameworthy gives us good reason to blame him, it doesn't settle the question of whether we ought to.[78] And Manny's case is a good application of this distinction. I think it is a good application because at least some of the reasons one tends to offer in defense of the claim that Manny isn't blame*worthy* are really considerations for why we shouldn't blame him. One natural consideration to note is that coercion of the kind Manny faces is very difficult to resist. In fact, we might go so far as to claim it would be unreasonable to expect Manny to refuse Moira's command. We think things like we would have chosen as Manny did in similar circumstances.[79] Neither

[76] I make similar claims about how our intuitions may be affected by claims regarding "shared" responsibility in Chapter 4, Section IV.3.

[77] See D'Arms and Jacobsen [2000]. I discussed this phenomenon in Chapter 1.

[78] Manuel Vargas distinguishes between an agent's being responsible and it being the case that we ought to *hold* him responsible. See Vargas [2004], p.225-226. I take this to be the more general form of the point above, but since excusing only arises in the face of potential blame, I stick to the negative side of the things here.

[79] Both Watson [1987] and Graham [2005] point to similar reasoning, though they apply it to different phenomena. Watson discusses our inclination to excuse (at least in part) vicious criminals with abused

of these thoughts, it seems to me, establishes that Manny isn't blameworthy. They seem to have the wrong emphasis. The first thought regards the reasonableness of our *expectations*, while the second concerns predictions of our action in similar circumstances. These thoughts seem to have more to do with the appropriateness of us actually blaming him, rather than the appropriateness of blaming, *simpliciter*. So we may very well think it would be inappropriate for *us* to blame Manny. It would be hypocritical to blame him for giving in to an incentive we ourselves could not resist. But this does not support the view that blame is therefore inappropriate. Nevertheless, such considerations can serve as decisive reasons for not actually blaming Manny.[80] We can take these as reasons that show why we shouldn't blame Manny, *despite* his blameworthiness. Conflation of these two issues serves to confuse judgments about Manny's blameworthiness, and the role of coercion in undermining responsibility.

I think, therefore, that we cannot rely merely on intuitions of Manny's blameworthiness. They are too much affected by irrelevant considerations. And the tough cases, ones in which Manny must take more life than he saves, are tough independently of whether or not Manny is responsible for the deaths. They are tough because our substantive moral theories lack clear answers in these cases, and because our tendency for shifting blame to the coercer and empathetic judgment clouds our assessments.

childhoods; Graham discusses our inclination to excuse on the grounds of moral ignorance. Graham explicitly rejects such reasoning, whereas Watson is more descriptive in his stance. I discuss these claims again and in greater detail in in Ch. 4, Sec. 3.

[80] Similarly, I think, we can defend not punishing Manny in similar ways. However, the issue of punishment, and its diverse goals, would complicate this discussion considerably. I will not discuss it further here.

3.3.3. Objection 2: Coercion and Praiseworthiness

A SIMILAR OBJECTION can be pushed if we consider cases of coerced good deeds. The objection claims that if coerced agents were responsible for the good they bring about, then, on my view, this would mean that they are praiseworthy for such things. But agents who are coerced into doing good are not praiseworthy, so there is something wrong with my account. Notice again that this is not an objection to coerced agents being responsible so much as a critique of my explanation of praiseworthiness. For even if agents weren't praiseworthy for coerced outcomes, this alone wouldn't show that they aren't responsible for them. Still, the objection does put pressure on my overall account, and so is worth addressing here.

First, let's consider a case or two. Suppose Moe holds a gun to Marlene's head and demands that she donate $10 to Oxfam. Supposing that Marlene donates the money, is she really praiseworthy for it? To ensure this is a case of coercion, let us further suppose that Marlene doesn't want to donate the money. While she can afford it, she would prefer to spend the money elsewhere. In fact, she would rather avoid either of the outcomes, just as in Manny and Micah's cases. She is also clearly presented with a powerful incentive for donating the money, since Moe will kill her if she doesn't. So, structurally, this seems to be a case of coercion. Marlene also seems to make the appropriate choice. And according to our characterization of coercion, it would seem that Marlene is responsible for the donation. Given my account of praiseworthiness as being responsible for something good, and given the fact that donating money to Oxfam is good, Marlene is praiseworthy for it.

But the objection claims this is the wrong result. One who voluntarily donates money to Oxfam may be praiseworthy, but this has to do with the concern for others that he must have, and Marlene lacks that concern since she is primarily motivated by the concern for her own life. That, the objection claims, is not the proper motivation for giving to charity. A similar sentiment, I take it, lies behind our reticence to praise an analogous agent. Consider the Calculating Benefactor. He is embarking on a new business venture soon. In order to garner positive publicity, he makes a substantial donation to a local charity. This donation, he knows, will reflect well on him in the media, and the exposure will help his business venture prosper. He does not, it turns out, care one whit about the people he's helping; his sole motive is his business venture's success. I take it that many people think that the Calculating Benefactor is not praiseworthy.[81] The thought is that he acts from the wrong sort of motive. His aims are not noble, he lacks the requisite concern and respect for those the charity serves, and thus is not worthy of praise for his donation. Call this the *Right-Motive Account* of praiseworthiness. According to the Right-Motive Account, an agent is only praiseworthy for an outcome if it is good, she is responsible for it, and she acts from the right motives in bringing it about. Such a view assumes that one's motives play a role in one's praiseworthiness for an action or outcome. But while I understand the intuition, I don't see why this should be the case. Similarly, I don't see why Marlene isn't praiseworthy for donating to Oxfam. After all, it is a good thing to donate money to charity, and she is responsible for her donation.

[81] Joshua Knobe has performed experiments in which an analogous case is described (although it concerns a side-effect of an action). A significant majority of respondents agree that the agent doesn't act intentionally (Knobe [2003], [2004]. In personal discussion, Knobe has said that similar experiments reveal the majority also think such agents aren't praiseworthy either.

I want to defend this claim by showing that its implications that naturally worry us (e.g., that Marlene is praiseworthy for donating to Oxfam) should not worry us. Motivating this defense is the observation that sometimes we speak of a person being praiseworthy for his character, and sometimes for a particular outcome. I think it is this distinction that can help explain the worrying intuitions. For example, the claim that Marlene and the Calculating Benefactor are both responsible for the outcomes they bring about does not entail that either one is a particularly good person. That is, the outcomes need not reflect strongly anything about their characters. In fact, we might think that since they do the right thing but not for the right reasons, *this* fact reflects poorly on their characters, in a way that overshadows whatever positive light being responsible for something good is able to shine on them. On this account, the quality of one's character is not determined solely by the outcomes one is responsible for. Rather, the quality of one's character may be a complicated weighing of the quality and strength of one's motives, intentions, dispositions to act, beliefs, and attitudes towards others, whether these mental states are ever acted upon or not.

On this account, then, we get the right contrast between the Calculating Benefactor and the Benevolent Benefactor. The latter donates the large sum to a charity because he is motivated by the desire to help others, he feels deeply the affected group's suffering and wants to help alleviate it. Surely he is a more praiseworthy individual. But here we get an explanation for our reluctance to praise the Calculating Benefactor and Marlene. They seem to lack praiseworthy *characters*. They only donate because of egoistic incentives, whereas the Benevolent Benefactor acts out of his concern for others. We needn't assume, however, that the connection between responsibility for outcomes and quality of characters is so tight. The fact that the Benevolent Benefactor has a more praiseworthy character, even that he is more praiseworthy *overall*, does not show that the Calculating Benefactor and Marlene are not each responsible for their donations. It just shows that they aren't particularly good people. And this fact may well be enough reason to not actually praise them for their actions. Just as in the case of coerced bad deeds, we can distinguish between the correctness of a praise attribution and whether or not we ought to actually praise someone. In these cases, we seem to have decisive reason not to praise them, but it does not follow that it would be a mistake to do so, so long as we restricted our praise to the good outcome, and not their characters.

As a final consideration, allow me two more comparisons. Compare the Calculating Benefactor to the Calculating Non-Benefactor. The latter is also planning to donate money solely to improve his public image and garner positive press for his new business venture. But after deliberation, he decides that the cost would not be worth the benefit, and so he refrains from donating any money to charity. Here, he suffers from the same wrong motivations as the Calculating Benefactor. But it seems right to suppose that the Calculating Benefactor is better than the Calculating Non-Benefactor *with respect to what each did*. At least the former actually benefited other human beings. He actually helped, whereas the Calculating Non-Benefactor had bad motives *and* failed to help anyone. Again, we can explain this difference if we accept that the Calculating Benefactor is responsible and praiseworthy for his donation.

Similarly, compare Marlene to Maxine. Maxine, too, is threatened with death if she doesn't donate $10 to Oxfam. She, too, would prefer to spend that money elsewhere, and just like Marlene, she would rather avoid either donating the money or getting shot.

But unlike Marlene, Maxine refuses to donate the money, and she is shot. While perhaps there is something to be said for standing up to coercers, it seems as if Maxine does worse than Marlene. At least Marlene ends up donating money to charity. Maxine not only shares Marlene's bad dispositions and motives, she still refuses to help others even when faced with death.[82] Again, we can explain why Marlene is better if we accept that she is responsible and praiseworthy for her donation. It might be suggested that Maxine's actions are evidence of a severe character deficiency. She is so bad that even under threat of death she refuses even to donate $10. This action seems to display an utter disregard and perhaps an active contempt for those suffering. We might then ask, couldn't this character deficiency explain why Maxine is worse than Marlene? And perhaps it can. But I don't need to eliminate all rival explanations in order to defeat the objection; I only need to show my account to be consistent with the phenomena. This is accomplished so long as appeal to Marlene's being praiseworthy for donating money can show her to have done better than Maxine, and to be, at least in that respect, better than Maxine.

In both cases, it may well be that the badness of the agent's motives and dispositions far outweighs the positive contribution being responsible for the donation confers. Moreover, being forced to do good may itself be a consideration that reflects poorly on an agent's character. Nevertheless, these considerations serve to show that the evaluation of an agent's character is affected by more than the outcomes for which the agent is responsible. And given these observations, we need not conclude that the agents

[82] We might question Maxine's rationality. Even on a modest cost/benefit analysis, surely one's life is worth more than $10 dollars, especially to oneself. If this poses a problem, then merely suppose that Maxine doesn't value her life more than, say, $5. While odd, this additional fact doesn't seem to change her moral evaluation as a person.

are not responsible, indeed praiseworthy, for the outcomes they bring about, even though they are coerced. Indeed, we can maintain that they *are* praiseworthy for these outcomes, and still explain the differences in the evaluation of their characters (i.e., their praiseworthiness 'as people') as a function of the multitude of considerations that weigh into such an evaluation. Doing so preserves a natural view about responsibility and praiseworthiness, one that does not seem affected by coercion in cases of blameworthiness, and can account for the apparent unintuitive results such a natural view gives about the cases above. It does so, I contend, because our intuitions in such cases can be explained primarily by facts about *character* not responsibility for *outcomes*.

3.3.4. Objection 3: Non-Coerced Agents

ANOTHER OBJECTION SEEKS to put pressure on my position by comparing coerced agents to non-coerced agents. Compare Otis, for example, to Oscar.[83] Oscar hates his neighbor, and one day, while his neighbor is away, Oscar burns down his field. Oscar is surely blameworthy for burning down his neighbor's field, and he is therefore responsible for it as well. I am committed to the claim that Otis is responsible and blameworthy too. I have argued that while we have decisive reasons for not actually blaming Otis, it wouldn't be a mistake to do so. But the objection notes that surely Otis is *less* blameworthy than Oscar. After all, Otis burns the field down because he has to if he is to save the town. It doesn't reflect poorly on his character at all; he did the right thing. Oscar, on the other hand, burns down the field to harm his neighbor, and it does reflect poorly on his character. He did something wrong. But if both are responsible for burning

[83] Here I am using an example of necessity, since I think coercion and necessity are of a kind. The same argument could be run with Manny, of course.

down a field, and I’ve claimed that burning down the field is bad, then they should be equally blameworthy.

In reply, let me again note that this objection again would not show that coercion isn’t a justification. Indeed, it seems that it would have to be a justification in order for Otis to have done the right thing. That is, it must be the case that something shows why burning down the field was the proper choice in the situation. Still, we might think that Otis is surely less blameworthy than Oscar, though both are supposedly responsible for burning down a field.

I agree that Otis is less blameworthy, in general. That is, I believe that if we are just comparing relative levels of blameworthiness between the two agents in their given scenarios, it is clear that Otis is less blameworthy than Oscar. But, I maintain, this has relatively little to do with burning down a field. Instead, I think what drives the judgment that Otis is less blameworthy is the fact that there is something Otis is praiseworthy for. After all, Otis saves the town from burning down. And this is a very good outcome. Indeed, it is so good, that it outweighs the bad caused by burning down his neighbor’s field. This is what a justification amounts to: an agent is justified in bringing about some bad outcome when greater good is achieved. More may need to be added to this formulation, but it gets the core of justifications right. For example, suppose Peter threatens to punch Paul if the latter doesn’t write a curse word on the side of the school. This seems to be a case in which Paul chooses well if he accedes to Peter’s demands. Creating a little graffiti is the better choice than suffering a broken nose.[84] But suppose that instead of graffiti, Peter’s demand is that Paul murder the principal, Pam. Here, I

[84] Though, again, the example is a bit contrived, since it is open to Paul to do something other than the two alternatives presented.

suspect, we think it would be improper for Paul to choose killing Pam over the threat of a broken nose. Obviously, physical harm is to be avoided, but the threat of being punched just doesn't carry enough weight to justifiably coerce when someone's life is on the line. So, if Paul does kill Pam, I suspect we think him both responsible and blameworthy for it. The differences between these two cases seem to be with respect to the appropriateness of the choice made given the relative incentives involved.

Similarly, we can agree that Manny chooses properly in cracking a safe over sacrificing his life. But suppose that he is threatened with a kick on the shin, or even a verbal insult, if he fails to crack the safe. Here it is not at all obvious that in choosing to crack the safe he chooses the proper course. So, even though both options are unattractive, coercion offers little defense if the agent chooses poorly, as Manny does if he opts for cracking the safe over being called a "scumbag."

So it seems that coercion only justifies the agent's choice when there is a significant enough threat to outweigh the reasons against doing the alternative in the first place. It is clear that whether or not a given choice is justified will depend on the options available to the agent. I'm not in a position to state what counts as an "available" option, but we can easily rule out the logically and physically impossible, and perhaps the psychologically impossible as well. I trust that we needn't worry over how to draw the appropriate line here, since it seems a rather straightforward matter for cases like Manny's. There, as I noted, he has the option of resisting Moira, of trying to get the gun from her, of running away, and many other possible choices. Moira's presentation of the options takes the form of a false dilemma. Manny isn't constrained by her dictates, but can opt for an option that's "not on the table."[85] But so long as he chooses appropriately, he is justified. And presumably choosing appropriately involves something like choosing an option reasonably as good as any alternative.

Justifications, then, are all things considered notions. They appeal to the conditions in which the action took place, especially the outcome that was avoided by taking the action, and that the action was the better option of those available. Blameworthiness perhaps can be an all things considered ascription. We might take two agents and ask, "Which is more blameworthy?" The question is unfortunately vague, but we might take it to mean, "Given all the things each is responsible for, which of the two is responsible for the worst stuff?" Or, to put it another way, we might be asking who's more blameworthy for their characters, and their attitudes, and their deeds, etc., aggregated over their lifetimes. I'm not sure how to settle such questions. Indeed, even limiting discussion to a much smaller portion of time, say, a particular sequence of events still leaves us with a difficult question if what we're after is overall blameworthiness for everything relevant to that sequence. I think it difficult to know how to handle ascriptions of overall blameworthiness precisely, since it isn't clear how to aggregate blameworthiness over a vast array of objects of responsibility.[86]

That being said, I think coercion and necessity mitigate overall blameworthiness because they stand as testimonies to the agent's motivation. Coercion and necessity are relevant because they give an explanation for why the agent acted contrary to important reasons against doing bad things: it was because something even worse would have occurred otherwise. Agents who act under such circumstances do not have bad motives

[85] Interestingly, Sartre makes a similar observation. See Sartre [1943, 2003], pp.255-281.
[86] This is a very interesting question I hope to pursue in later work.

as a result.[87] On the contrary, we might think they chose as they ought to have done. And yet the features that make the chosen option bad remain even when it is the proper choice. After all, the trite phrasing of such scenarios is "the lesser of two evils," not "what was once an evil but is no longer so given the possibility of greater evil." So I think there is some blameworthiness for doing bad things even when they are justified, and even though we shouldn't actually blame the coerced.

3.3.5. Objection 4: Responsibility and Character

STILL ANOTHER OBJECTION goes as follows. Responsibility is supposed to be the sort of notion that connects actions with an agent's character. An agent's being blameworthy for x reflects negatively on his character because of the negative nature of x (i.e., that it is bad). Being praiseworthy reflects positively. But for Manny or Otis, this isn't true. So it is mistaken to think they are responsible. Manny's character isn't negatively impacted by his cracking the safe, nor is Otis' character for burning the field. And yet these are bad things, and so if they were responsible for them, their characters would be negatively impacted as a result. Therefore, they must not be responsible for these things.

I agree that when an agent is responsible for an outcome, this tells us something about the agent in light of the nature of that outcome. And I agree that in cases of coercion and necessity, what the nature of the outcome tells us is different than in ordinary cases of action. But the objection has the wrong view of how my account treats these cases. Justifications are cases in which bad outcomes are relativized in a certain

[87] Or, at least, we cannot infer that they do from their "bad" action.

way due to the surrounding circumstances. That's why Otis' being responsible for burning down the field isn't quite the same as Oscar's.

Part of the explanation for the difference concerns the sense in which Otis' action isn't divorceable from the particular circumstances in which it is performed. It is these very circumstances which allow his burning the field to be justified. The explanation of his action includes the beliefs featuring in Otis' reasoning, at least some of which concern performing the act *so as to avoid* a substantially worse harm. When Oscar acts, of course, he has no such beliefs. When Otis is responsible for burning down the field, he is justified so long as he does it to thereby save the town,[88] which means that he is also responsible for saving the town. So, while in cases of justification we will be able to separate conceptually the agent's responsibility for the bad outcome and his responsibility for the good outcome, the same action precipitates both. We will not, therefore, be able to pry what his responsibility for burning the field tells us about him from what his responsibility for saving the town tells us. For the responsibility relation seems to tie Otis to both outcomes simultaneously, since the same action precipitates both, and he intentionally brings about the one outcome in order to bring about the other. And this, I submit, is the unique feature of justifications. His blameworthiness for burning the field down is clearly outweighed by his praiseworthiness for saving the town, and this will be true for all justifications, for they are justifications precisely *because* they are instances in which the agent does something bad that nevertheless averts something worse. And averting the worse outcome will carry with it the relevant praiseworthiness outweighing the blameworthiness from the bad outcome in something like a proportional

[88] We might include a qualification that it has to be true that burning the field will save the town. This would depend on whether an agent's beliefs alone can justify, or whether the beliefs have to be true, as well, or something in the middle (e.g., the beliefs must meet some reasonability standard).

fashion. The greater the harm avoided, the more praiseworthiness to outweigh blameworthiness, and thus, the more justified one is in performing a harm.

This account also nicely fits phenomena surrounding particularly difficult choices where it might not be clear which option is preferable. In such scenarios, I take it, the relative badness of the options available will be much closer, and thus the praiseworthiness corresponding to the outcome averted will, perhaps, only slightly outweigh that accrued by the action performed. Thus, in such scenarios, it will be easier to see how the agent might be blameworthy for the bad outcome he brought about, even if in the end we agree that he ought not be blamed for it. What helps explain this feature of justifications generally is that the fact that his praiseworthiness for averting some terrible outcome outweighs his blameworthiness for the bad outcome he does bring about gives us good reason not to actually blame him. If we did, we would be unfairly ignoring the praise he is due for averting the terrible outcome.

Now it may also be the case that he ought not actually be praised for averting the terrible outcome. I take it in many (perhaps all) cases of coercion and at least some cases of necessity, this is the case. I suspect that this is true in cases where the respective bad outcomes are closer in degree, so that praising the agent for adverting disaster would be to ignore the reality of the bad he brought about in order to do so. Again, dilemmatic cases will be good examples, though I think this can be true even of cases in which there is a clear enough answer as to what the agent ought to do. Moreover, in cases in which we would likely praise the agent, I think this most naturally reflects a commitment on our part to reinforce choosing correctly. So, for example, in Otis' case, while there seems to me something odd about praising him for saving the town, since this ignores the fact that he burned down his neighbor’s field to do so, there may be good prudential reasons to praise him anyhow, if only to foster in others the tendency to overcome the strong reasons against bad actions necessary to prevent even more serious harms.

3.3.6. Objection 5: Coercion and Freedom

I CONSIDER NOW A FINAL objection that has a quite different source. There is a general view, proposed by Gideon Yaffe, that claims coercion intuitively restricts an agent’s responsibility-relevant freedom, and thus undermines his responsibility. The idea here is that responsibility requires a kind of freedom, and that freedom is impaired by coercion. Thus, responsibility is undermined as a result.[89] Coercion could restrict freedom, it is claimed, by limiting the available options one can choose from. If we imagine a range of scenarios in which a coerced agent could choose, coercion is characterized by the coercer adjusting tactics so as to guarantee (as much as possible) compliance with whatever he wants done. Thus, no matter what sorts of reasons Manny might act on, Moira will attempt to insure that she provides the strongest reasons in favor of cracking the safe in all cases. That is, if Manny’s choosing based on reasons of self-interest alone, she’ll threaten his welfare. If he’s privileging other-interested reasons, then she’ll threaten his family, and so on. The objection claims that coercion makes the array of reasons-responsive mechanisms that Manny might act on “functionally equivalent”[90] – that is, no matter which mechanism is engaged, the outcome Moira wants will result.

I have two replies. The first is that this picture of coercion is overly restrictive. It unnecessarily assumes a particular view about freedom, one that doesn’t seem to me to be

[89] For a presentation of such a view, see Yaffe [2003], esp. pp.350-355.

[90] Yaffe [2003], p.353.

plausible enough to assume. Second, even granting its conclusion, it fails to establish that responsibility is undermined. I'll take these replies up in turn.

First, the intuitive picture is that coercers restrict the available options open to the coerced. Manny would prefer to do a range of other things, but Moira is poised to present him with incredibly strong incentives at every turn. The picture suggests that Moira makes it such that given the various ways Manny might choose, the result will always be the same – he will crack the safe. His freedom with respect to what he does is restricted by her manipulation of his reasons for choosing.

But this picture is infelicitous. Suppose, for the moment, that the thesis of determinism is true. Suppose it is true that every event e after some time t is entailed given the facts of the world at time t and the laws of nature. If this is so, then there aren't any options open to Manny. Indeed, one way to state the thesis of determinism is that "there is at any instant exactly one possible future."[91] What Manny does is determined, just as what Moira does is. His available "options" are already set; this is what determinism holds. The picture of Moira's interference sketched by the objection suggests that she is interfering with Manny's choosing processes, such that the result will always be the same. But given the truth of determinism, Manny and Moira are themselves parts of a deterministic process, which fixes what each will do. So it is misleading to suggest that Manny's freedom is limited by Moira anymore so than it is by determinism itself. It follows, then, that the objection only establishes its conclusion if determinism is false. But that seems to me an unreasonable and implausible assumption to make, and it is certainly one to be resisted by compatibilists, especially if there's an

[91] Van Inwagen [1983], p.3.

alternative account of coercion on offer.[92] Yaffe does suggest that determinism itself is different from a coercer's actions, for the laws of nature will not change tacks to facilitate the compliance of the coerced, whereas Moira will presumably adjust tactics to ensure compliance. But if determinism is true, we cannot, it would seem, divorce Moira's actions from the total determined system, and her changing tacks would be no more or less determined than a ship's captain being "coerced" by a storm to dump his cargo overboard (an example of necessity).[93] So while this suggestion does seem to point to the coercer's shared responsibility for actions, I do not see how it effectively contrasts coercion from necessity.

Second, even if we grant the conclusion, those pressing the freedom objection must still show that a restriction of freedom implies a reduction in responsibility. In other words, proponents must show that maximal freedom is necessary for responsibility. Even were it the case that Manny is less free in the face of Moira's coercion than he would be otherwise, this is insufficient for demonstrating that his responsibility is similarly reduced. Perhaps there is merely a threshold level of freedom required for responsibility, so that one could fail to be maximally free and still be maximally responsible. I do not claim this is necessarily so, only that its possibility requires refutation. At the very least one cannot simply assume that freedom and responsibility are directly proportional, as the objection seems to do.[94]

[92] It should be noted that Yaffe isn't trying to give an independent account of coercion. Rather, he's assuming that coercion undermines freedom and trying to show how that could be the case. In my mind, he does so only by eliminating compatibilist accounts of freedom, and by extension, compatibilist accounts of responsibility.

[93] A contrasting example Yaffe uses, borrowing from Aristotle.

[94] Yaffe certainly seems to make this assumption. See Yaffe [2003], p.335, n.1.

In light of the above, I conclude that coercion and necessity are of a kind, that *neither* undermine responsibility, and that therefore agents who act under coercion or necessity are responsible (*ceteris paribus*)[95] for what they bring about. Coercion and necessity are both justifications, and so, as long as the agent chooses properly, he ought not be blamed or praised for the outcome, even if he can still be blameworthy or praiseworthy for them. I think this account explains the phenomena involved and maintains the natural view that blameworthiness and praiseworthiness are just responsibility for bad and good things, respectively.

3.3.7. A Brief Note on Compulsion

VERY BRIEFLY, THEN, I WANT to make a final observation. If I am right in my characterization of coercion and necessity, that they are defined in part by the presentation of a powerful incentive for action, then it may very well be the case that compulsion ought to be treated similarly. If klepto- and pyromaniacs can resist their psychological urges (i.e., if such individuals do not *always* act on such dispositions), then it would appear that such cases are structurally similar to cases of necessity. The agents in compulsion cases are similarly presented with a powerful psychological incentive, this time from "within," to bring about a particular outcome.[96] But if coercion and necessity fail to undermine responsibility, then I conclude that compulsion will similarly fail. However, unlike coercion and necessity, compulsion does not seem to be a justification;

[95] The qualification is necessary since one might be coerced to do something and still bring about an outcome accidentally. It would seem that responsibility for such a side-effect is still undermined in this case.

[96] The claim here is clearly conditional on psychology and neuroscience giving us the correct picture of such disorders. If, say, kleptomania doesn't work as I've described, then unfortunately my comments on coercion will not apply to kleptomaniacs.

it does not show that the outcome chosen was good or proper. As a result, if compulsion does not undermine responsibility, than klepto- and pyromaniacs are responsible for the outcomes they bring about (*ceteris paribus*), and since those outcomes are usually bad, they are blameworthy for them.

Similarly, and I think tellingly, we don't think those compelled to do good things should be less responsible. As Susan Wolf has shown,[97] if someone is presented with a powerful psychological incentive to donate to charity, or help a drowning swimmer, we aren't inclined to think their responsibility is undermined. Indeed, they may be praiseworthy. My account treats both types of cases identically, holding that responsibility is preserved in both. To my mind, this is a welcome result for such cases.

3.4. Conditions on Being Responsible

IN THIS SECTION, I WANT to review what has been said so far and outline the conditions on being responsible suggested by the characterizations of the classes of undermining factors provided above. So far, we have characterized three classes of undermining factors.[98] First, there are those considerations that show the outcome was the result of involuntary behavior.[99] These are considerations like spasm, reflex, and incapacitation. This class is The Involuntary. Second, there are those considerations that show the outcome was an unforeseen side-effect of action. These are considerations like accident, inadvertence, and ignorance. This class is The Unintentional. Finally, there are those considerations that show the agent acted from false beliefs about what she was doing. These are

[97] See Wolf [1980]. I discuss her conclusions in Chapter 1 as well.

[98] Recall that considerations of 'coercion' have been excluded.

[99] The qualification is necessary as many wouldn't want to count these sorts of bodily movements as actions. I remain agnostic on how we ought to classify actions for the purposes of action theory.

considerations of mistake. This class is The Mistaken. Examination of this final class also sheds light on why mistaken beliefs are relevant. They help specify the intentional structure of the agent and how mistaken beliefs can disrupt that structure.

My methodology was to presume that the undermining factors could operate by showing that a condition necessary for responsibility wasn't present. This presumption was supported, in part, by the observation that the undermining factors seem to be organized around presenting negative features of the outcome or agent; that the outcome was *un*intentional or *in*voluntary, or a belief was *mis*taken. The suggestion was that these negative features would themselves provide the conditions on being responsible.

In light of the above discussion, I think there are 3 conditions on responsibility for outcomes. In general, I'm concerned with the things that agents bring about. These can be talked about roughly as actions, events, or states of affairs. Agents sometimes aim to bring certain "ends" about, and these outcomes also sometimes have accompanying side-effects (or the actions that bring them about do). I intend to be neutral across possible questions regarding the individuation of actions, events, and the like.

So, for an agent to be responsible for an outcome, that outcome must have been (1) brought about voluntarily, (2) brought about intentionally, and (3) brought about without mistake. I need to say more about how these conditions relate to each other and what they rule out. I shall take each one up in turn.

3.4.1. Voluntariness Condition

AN AGENT BRINGS SOMETHING about voluntarily if it is the product of some action of his. As I understand it, when we explain the actions of agents, it often suffices to do so by reference to a belief-desire pair (or a set of beliefs and a set of desires). Stan gets up and goes into the kitchen and gets a soda from the fridge. Why did he do that? Well, we might say, Stan wanted to drink a soda and he believed there was a soda in the fridge. So, he went to the fridge and got one. I take it all actions for which one can be responsible can be explained in this way, whereas not all events can be so explained (e.g., those events that are not also actions, like the eruption of a volcano).[100] An agent satisfies the voluntariness condition so long as the outcome was produced by an action that can be explained by reference to a belief-desire pair.

This construal of the condition rules out just those outcomes that gave rise to the condition in the first place. For example, when Barbie suffers a seizure and hits Ken in her flailings, the outcome is explained solely in terms of her seizure. She doesn't "act" at all, in the sense made relevant by belief-desire pairs. It is important to note here, however, that it is not the fact that we can explain the outcome in non-belief-desire terms that violates the condition. This may be possible in even paradigmatic cases of intentional action.[101] The relevant fact is that we *cannot* explain the outcome in belief-desire pairs. Similarly, if Adam breaks a window while sleepwalking, or Barry breaks a window while swatting at a swarm of bees, they too do not act from a belief-desire pair. As a result, none of these outcomes is voluntarily produced, and thus each fails the voluntariness condition. Responsibility for each is undermined.

[100] Also, there may be some actions that cannot be explained this way. On some views about action, distinguishing between actions and mere events can be done roughly by distinguishing between the "stuff people do" and the "stuff that happens to them". For an example, see Brand [1984], esp. pp. 3-6. I remain neutral on whether the concept of action is best thought of in these terms or the ones outlined in the body of the chapter. I do think, however, that belief-desire pairs are necessary for an action to be a proper object of responsibility, and I will call all such explainable objects as "actions" simpliciter. I say more about this condition in Chapter 5.

[101] Indeed, I should think this necessary if physicalism about the mind is true.

3.4.2. Intentionality Condition

THE INTENTIONALITY CONDITION constrains what must be true about the belief figuring in an explanation of the agent's action. An agent satisfies the Intentionality Condition so long as he believed that the outcome in question might occur as a result of his action. What you intend to do is generally susceptible to what you believe you can do. But for the purposes of responsibility, it isn't necessary that the outcome be intended. Instead, one need only foresee that the outcome may occur. Moreover, the agent need not know that the outcome will occur, or even that it will likely occur. Rather, it is sufficient for intentionality (as responsibility requires it) that the agent foresees the sheer possibility of the outcome in question.[102]

Again, the construal of this condition rules out just those outcomes that initially gave rise to the condition. When Fred shoots Barney in the eye, he doesn't believe that the injury is a possible result of his opening the bottle of champagne.[103] And when Lenny steps on his friend's hand inadvertently, he also doesn't believe that the result is a potential consequence of his stepping. Indeed, if he did, he likely wouldn't step. And when Martin opens the door and hits Mark in the face, he also doesn't believe that opening the door will cause injury.[104] Again, if he had such knowledge, he would likely act differently.[105]

[102] The condition thus adopts a wider scope than might be suggested by certain uses of the term 'intentional.'

[103] By stipulation. Indeed, it doesn't even seem to be a "foreseeable" result. Rarely, if ever, do champagne corks injure people.

[104] Here I seem to join those like Galen Strawson and Fischer & Ravizza who claim there are subjective conditions on responsibility. What the agent himself believes is relevant.

[105] This counterfactual tests are meant to illustrate the point, not as elucidations of how the condition must be satisfied. One could satisfy the condition, and bring something about unintentionally, even if counterfactually he still would have acted as he did. In the counterfactual scenario, however, he would no longer bring the thing about unintentionally.

One might accept the foreseeing aspect of the condition, and yet be skeptical about the possibility aspect. One might object that it surely matters how likely it is that a particular outcome will result from a particular course of action. Allow me to comment briefly on why this isn't the case.

First, notice that in cases of direct intention, the chance of success seems immaterial to ascriptions of responsibility (and, indeed, intentionality).[106] For example, suppose Leona is trying to shoot Lionel. But she's never shot a gun before. She's quite a distance away, in the wind and rain, and has trouble aiming the gun. Let's assume that the chances that she successfully hits Lionel are incredibly low. Indeed, we can set her probability of success arbitrarily low; say, she has a one in a million chance, literally.[107] Nevertheless, if Leona successfully shoots Lionel, she does so intentionally and she is responsible for shooting him. Success, no matter how improbable, is all that's required for meeting the intentionality requirement in cases of direct intention.[108]

But the objector likely has in mind cases of indirect intention. That is, the objector is claiming that in cases where harm is merely foreseen as a possible side-effect of some course of action, then the likelihood of that harm actually occurring does become relevant to ascriptions of responsibility. For example, suppose Harry wants to put down some pesticide at the edge of his property. He wants to do it today, on Sunday, because it's his day off. He knows, however, that if it were to rain within 4 hours of putting the pesticide down, the rain will wash some of the pesticide into his neighbor's prized

[106] Remember that, when I speak of intentionality (and intentionally x-ing, say), it is always qualified with "as responsibility is interested in the notion". Some may want to refine the notion of intentionality for technical reasons involved in other debates. I'm only interested in specifying the sort of mental states required in order to be responsible for outcomes.

[107] If this probability seems too high, the reader is invited to take me seriously, and indeed set the probability as low as he or she likes. I trust judgments about the case will remain unchanged.

[108] In cases of direct intention there is necessarily an explanatory belief about the potential outcome: it is the very aim of the action.

petunia patch, killing them. Now suppose that Harry knows there is an 80% chance of rain for the next 4 hours. If he puts down the pesticide, and it rains, and his neighbor's petunias are killed as a result, it sure seems as though he kills the petunias knowingly (which suffices for meeting the intentionality condition). That is, he believes that killing the petunias is a potential result of his action. And it surely seems as though Harry is responsible for killing them.

The objector will try to put pressure on this claim by reducing the probability of rain. So, suppose that everything is as before, only now Harry knows that there is only a 5% chance of rain for the next four hours. If he puts down the pesticide, and it rains, and the petunias are killed, I still think he kills them knowingly and is responsible. Indeed, so long as Harry believes that his putting down the pesticide may lead to killing the petunias, and his putting down the pesticide does lead to killing the petunias, he does so knowingly and is responsible for killing them.

To make the above claim more plausible, notice what happens if we ratchet up the seriousness of the harm risked. Killing petunias may be a bad thing, but it isn't that bad. So suppose that Vincent loves watching his favorite movie, "Field of Dreams." But he also knows that there is a 1 in a billion chance that if he puts the DVD in his player and pushes "play" everyone else on the planet will die. This is a miniscule risk (and we can reduce it further to any arbitrary amount). Nevertheless, should Vincent play his movie, and everyone else dies as a result, I think it clear that Vincent kills them knowingly and is responsible.[109]

[109] To keep things simple, let's assume the explanation for why pressing the button will kill so many people has nothing to do with any third-party's nefarious schemes or intervention.

There is a certain sort of strictness involved here. To put it one way, I'm claiming that knowledge of the possibility of an outcome can make one responsible for it. Acting in the knowledge that your action risks an outcome means that you "accept"[110] that outcome's obtaining. And the condition would seem to expand the realm of the things we can be responsible for. It is this expansion, I think, that lies at the heart of the objector's worry. I can believe a good many things are possible consequences of my action, but surely it is too much to demand that I am responsible for all of them that may occur. Indeed, the worry may be simply that "possibility" opens up too much; if the probability is allowed to be arbitrarily low, then certainly there are a great many things that are only barely possible, even restricting ourselves to physical possibility. We are all familiar with the image of a butterfly's flapping causing a hurricane on the other side of the world. Taking a liberty or two, if I am that butterfly, the objector asks, am I responsible for the hurricane?

This worry is unfounded. Remember, the Intentionality Condition is only strict when it comes to outcomes the agent believed were possible consequences. So to answer the objector, you the butterfly are responsible only if you knew the hurricane was a possible result. After all, it is surely possible that opening a champagne bottle can lead to the cork hitting someone in the eye. It is likely a rare occurrence, but not impossible. But Fred isn't responsible for hitting Barney in the eye. Why not? Because Fred doesn't believe that Barney's injury is a potential side-effect. And it is this consideration that undermines Fred's responsibility for hitting Barney in the eye. We're concerned with the beliefs Fred actually holds at the time of his action, not the facts about possible

[110] We must speak metaphorically here, since one need not endorse the outcome to knowingly bring it about.

outcomes.[111] Contrast this with a counterfactual case, where Fred thinks to himself, "Opening the bottle in this fashion might hit someone in the eye." When he hits Barney in the eye with the cork, it is far less plausible to suppose he's still not responsible for doing so.

3.4.3. No-Mistake Condition[112]

SO FAR, WE'VE SEEN THAT the Voluntariness Condition and the Intentionality Condition are satisfied when (1) an agent brings something about through an action that can be explained by a belief-desire set, and (2) that set includes a belief that the outcome in question was at least a possible result of the action. There is only one more condition necessary for responsibility. An agent brings something about without mistake so long as he has only correct beliefs about the nature of what he is bringing about.

Once again, this condition rules out mistakes. Recall our prime example. Jan takes what she believes to be her coat from a pile of coats at a party. She wants to leave and it is cold outside, so she wants her coat, and she believes her coat to be the red one now in her hands. But she's wrong. The coat she holds is someone else's, though it is the same type, color, and size. In other words, she has a false explanatory belief (it is doubly explanatory since it helps explains both why Jan took someone else's coat and why she didn't take her own). This false belief constitutes Jan's mistake; she was wrong

[111] There is a related objection that claims this condition may rule too much out at times. For instance, what if someone's actual beliefs are unreasonable? Suppose someone claims he didn't believe that shooting his gun haphazardly could lead to serious injury? I think this worry, too, is unfounded, but discussion of the relevant points is taken up in Chapter 3.

[112] I've previously claimed that one arrives at these conditions be negating the negations suggested by the undermining factors (the *in*voluntary; the *un*intentional; and the *mis*taken). For the other two conditions, the double negation invites a return to the simpler positive formulation (e.g., the voluntary; the intentional). Dropping the "mis-" from "mistake" does not invite such a return, as we don't have the appropriate corresponding English word. As a result, I'm forced to use the much more cumbersome locution in the text.

about the nature of what she was doing. She thought she was taking her own coat and not someone else's. Jan has a false belief about what she was bringing about, and this false belief seems to undermine her responsibility.

As stated, however, this condition clearly rules out too much. Not just any false belief about what one is doing can undermine responsibility. To use a modified version of an example from Pete Graham,[113] suppose Ben hates the Amish. He's a technophile and resentful of what he perceives to be an Amish disdain for technology and those who use it. While out walking one day he sees Jeb, who Ben takes to be an Amish person and proceeds to beat mercilessly. Now, Ben believes that he is beating a man and that the man is Amish. But Ben is wrong. Jeb isn't Amish at all; he's a Mennonite. So Ben has a false belief about what he is doing. He's mistaken about Jeb's religion. So he thinks he's beating up an Amish man, when really he's beating up a Mennonite. But surely this doesn't excuse Ben. He must be fully responsible for Jeb's injuries, if anyone is responsible for anything. So we must somehow restrict the sort of beliefs relevant to undermining responsibility, or else we risk excusing too much.

I think we do better to state the condition as follows: an agent brings something about without mistake so long as he has only correct beliefs about the nature of what he is bringing about necessary to generate an evaluation according to the appropriate normative standards. Allow me to explain.

Recall that, on my view, an agent is morally blameworthy[114] just in case he is responsible for something that is morally bad. He is responsible just in case he meets my three conditions. The "something" is morally bad just in case our correct moral theory

[113] Graham [ms 1], p.20.

[114] All that I say here can be said *mutatis mutandi* for praiseworthiness and all non-moral types of each.

says so. We care about responsibility principally because of the important role it plays in ascriptions of blameworthiness and praiseworthiness (across normative standards). As a result, the conditions on responsibility are sensitive to that role it plays. So, I think the relevant set of beliefs here is the one concerning those facts necessary for generating an evaluation in the circumstances according to our correct moral theory.[115] Let's take this version of the condition back to Ben's case.

To answer the question of whether Ben's beliefs about Jeb's religious affiliation are relevant to responsibility, I suggest we need only answer whether or not Jeb's religious affiliation would be relevant to the moral evaluation of the act, given a plausible moral theory. While surely some moral theories might take such a fact into consideration, I think we would have especially strong independent reasons to reject such theories. Instead, it seems to me, our most plausible moral theories, whatever they say, will treat religious affiliation as irrelevant for evaluating instances of physical assault. If the fact of Jeb's religious affiliation is irrelevant according to the normative standards, then I think Ben's belief regarding Jeb's religious affiliation is similarly irrelevant for responsibility.

Turning back to Jan, we can see that this version of the condition preserves our judgments in her case. For it seems obvious that facts concerning whose coat a particular coat is will be relevant for evaluating instances of coat-taking. And if that is the case, then I think beliefs about such facts are of supreme importance for responsibility. Thus, the restricted revision of the condition serves to segregate just those cases in which false

[115] For further discussion of this condition, see the discussion on "Normative Competence" in Chapter 4, Sec. IV.2.

beliefs undermine responsibility, and those in which they don't. Jan's responsibility is therefore undermined, but Ben's is not. And this is how it should be.

A final word on this last condition. In Chapter 2, I argued for a notion of responsibility that is significantly independent of our practices. And here, it may seem as though I am smuggling our evaluative practices back into that notion. But I'm not. I don't think the practices are constitutive of the notion of responsibility, nor do they explain what it is to be responsible. That was the heart of my critique in Chapter 2. Moreover, even here, I do not think that the evaluative practices play a crucial role. What we care about is that the agent's beliefs about the features of the case are correct. Some such features will be irrelevant in assigning blame (e.g., Ben's false beliefs about Jeb's religious affiliation), while others will not (e.g., Jan's false beliefs about the coat she's taking). And these are important for the purposes of responsibility, only because they indicate the level at which the agent is aware of what he is doing. A certain level of awareness is necessary (e.g., that one is hitting a human being and not a tree stump), while further awareness is not necessary (e.g., that one is hitting an Amish person rather than a Mennonite). The agent must be aware of those facts that help explain the moral verdicts, but he need not be aware of the moral verdicts themselves. He needs to know he's causing harm, not that causing harm is wrong. So the evaluative practices themselves are not important, only awareness of the facts relevant to those practices.

This level of awareness, I think, is not restricted to the moral cases either. I think in general the same level of awareness, or awareness of the same sort of features, will be relevant across normative domains. That one is using blue paint and not green paint is an analogous example from the aesthetic domain. Obviously, the color of paint is aesthetically relevant, and not morally relevant. So it isn't the case that the very same facts will be relevant across domains. But neither should we expect them to be. Just as the color of paint is morally irrelevant, so is the causing of harm aesthetically irrelevant. Different norms care about different facts. We might worry, therefore, that an agent could be responsible for the very same action under one set of norms and not another.

But far from being worried by such a conclusion, I think it the right result. It amounts to the claim that one could be morally mistaken while being aesthetically unmistaken. And I think this is correct. Imagine that some musician wants a particular sound on a given track. The sound he's looking for is a scream of pain. To get it, he plugs in his keyboard, sets the sound bank to "screams" and plays. Let us assume that the effect truly is aesthetically remarkable. It seems the musician is surely responsible for that quality, if anyone ever is. Now let us suppose that, unbeknownst to our musician, the keyboard works in the following way. Whenever he sets it to "screams" and plays a key, someone is tortured in order to elicit the proper tone. The technology is such that there is no lag time, and there was nothing in the materials that came with the keyboard suggesting this is how the sounds get created. I think our musician is not morally blameworthy for creating the effect. He doesn't know (and has no reason to believe) that he is causing pain. He believes sincerely that the screams are digitally created and no one was ever harmed in their creation. His mistake, I think, undermines his responsibility for causing the pain. And yet, despite his false beliefs, I think he is responsible for the remarkable effect he brings about in the song, and so he is aesthetically (or, musically) praiseworthy for it. For the aesthetic norms are not sensitive to the sources of the sounds, they care primarily about the quality of the sounds and their relation to the rest of the piece.[116]

So, one can be both responsible and not responsible for particular things in a given bit of behavior. In recording the screams he is both responsible for their aesthetic effect on the song and *not* responsible for the pain he causes in eliciting those sounds. I see nothing problematic with such a result.

It should be apparent from the preceding discussion that these three conditions fall into a hierarchy. Or, if one prefers, they increasingly limit what counts as a potential object of responsibility. The Voluntariness Condition limits us to actions (or the results of such actions) that can be explained by a belief-desire set. The Intentionality Condition limits us to those outcomes the agent believed (at least) might result from his action. And the No-Mistake Condition limits us to those outcomes in which the agent had all correct beliefs concerning those features of what he was doing necessary for generating an evaluation according the appropriate standards given the circumstances. The restricted set for each condition is an obvious subset of the previous condition's restricted set. Thus, all outcomes that satisfy the Intentionality Condition satisfy the Voluntariness Condition, but only some of them satisfy the No-Mistake Condition.

3.5. Conclusion

I CONCLUDE THIS CHAPTER by outlining the structure of blameworthiness and praiseworthiness, especially as it relates to the discussions of mistake and coercion. Recall that on my view one is blameworthy just in case one is responsible for something

[116] This is a claim that risks attack by seasoned aestheticians. I run the risk knowingly.

bad. And in many of the cases in question, we're interested in moral blameworthiness. For one to be morally blameworthy for something, she must be responsible for something morally bad. The 'for something' refers to an outcome, the principle object of responsibility with which I'm interested. Take Jan's case for instance. Now, it seems Jan is not blameworthy for taking the coat. And it also seems it is her mistaken belief that the coat belongs to her that excuses her from blame. Since such excuses show that the agent isn't responsible, my characterization of mistake suggests that responsibility for outcomes is concerned with outcomes under descriptions. This is because the outcome in question must be bad in order for us to be interested in an ascription of blameworthiness to Jan, so this means the outcome in question must be treated as 'the taking of someone *else's* coat'. For this outcome, Jan is not responsible, and her mistake shows us why. She took herself to be doing something else, namely, taking her own coat.[117]

But we also saw that coercion doesn't undermine responsibility, and that Manny is responsible for cracking the safe. Now, it *seems* that Manny also is not morally blameworthy for cracking the safe (even though I've argued that he is in fact). The defense of this claim was that Manny has a justification, namely, the coercion he acted under. Manny chose appropriately given the situation, so while he was responsible for cracking the safe, under the circumstances cracking the safe was the appropriate thing to do. Its appropriateness doesn't make cracking the safe itself a good thing to do. Rather, a justification notes that there was another outcome the agent is responsible for, an outcome that is good enough to outweigh the badness of the action coerced. In Manny's case, avoiding death was a significant enough good outcome such that it justifies bringing

[117] This is not to suggest that her taking of 'this' coat and her taking of someone else's coat weren't the same event, only that her beliefs regarding the second description are relevant because they were false.

about a certain amount of bad, namely, cracking the safe. Similarly, Otis, who burns his neighbor's field to save the town, also does something bad (burn his neighbors field), but this action is outweighed by the goodness of saving the town. In both case, the agent has a justification only if the goodness achieved is better than the badness brought about. Moreover, we saw that the justifications are such that we cannot divorce the agent's performance of the bad action from the good outcome. In cracking the safe Manny saves his life. In burning the field, Otis saves the town. Thus, overall, justified agents do not appear blameworthy because they are responsible for both outcomes, and because it is a case of justification, the goodness outweighs the badness, so we are apt to disregard in large measure the blameworthiness.

These two ways of mitigating blameworthiness fall out of the *structure* of blameworthiness. One is blameworthy just in case one is responsible for something bad. One can undermine blameworthiness by showing either (1) that one isn't responsible for the outcome; or, (2) that the sum-outcome in question isn't bad. The first disjunct corresponds to excuses (as undermining factors);[118] the second corresponds to justifications. Since ascriptions of blameworthiness involve both a responsibility component and an evaluative component, undermining either of these components undermines the overall ascription. Justifications are then special cases, for strictly speaking, the agent is responsible for all that he did, but the nature of justifications are such that we don't make an ascription of blameworthiness because the overall quality of the objects of responsibility is positive. The responsibility component is a relation between the agent, her mental states, and the outcome. The evaluative component is a

[118] This is again my technical use of 'excuse'; recall that sometimes we call considerations 'excuses' that just point to reasons one shouldn't blame.

measure of the outcome according to the relevant norms. Blameworthiness depends on the combination of these two components; an agent is blameworthy only when the responsibility relation is present (i.e., the conditions on being responsible are met) and the evaluative component is negative. Praiseworthiness has the same responsibility component, and the evaluative component is positive.[119] And when the evaluative component registers a null value, we have "neutral" responsibility. This is the sense, agreed on by many, that an agent can be responsible for morally neutral outcomes.[120]

On my view, responsibility for outcomes is a relation between an agent and an outcome characterized under the proper description, one that depends on the agent's mental states. Blameworthiness and praiseworthiness are functions of responsibility relations, such that when the outcome is bad one is blameworthy, and when the outcome is good one is praiseworthy. Given this two-component structure, blameworthiness and praiseworthiness can themselves be undermined even when the responsibility relation is left unfazed. This, I argued in Section 3, is why we ought to exclude considerations of coercion from our list of undermining factors. Coercion, like considerations of necessity, mitigated blameworthiness by showing that the outcome one is responsible for is good under the circumstances (all things considered), and thus we ought not blame the

[119] Similarly, one can undermine ascriptions of praiseworthiness by showing either (1) that one isn't responsible for the outcome; or, (2) that the sum-outcome isn't good. The considerations relevant here are not as popularly discussed as excuses and justifications, but as Chapter 1 demonstrated, the considerations that show (1) are the same in both cases. I also think that moral theory should be concerned with (2), though historically it has been in the business of moral principles and obligations, and their violations.

[120] In most discussions the claim is that agents can be *morally* responsible for (morally) neutral actions. The mistake in such discussions is supposing that there is a special *sort* of responsibility that is moral in nature. Recall my critique of such a view from Chapter 1. On my account, it makes all the more sense how one can be responsible for evaluatively neutral objects. In fact, it is compatible with neutrality across evaluative domains (e.g., an outcome's being artistically neutral, or athletically neutral). Thus, moral neutrality is put on a par with similar notions in the other normative domains.

agent.[121] Moreover, the two-component model is a direct result of the Simple Solution to the Symmetry Challenge from Chapter 2. The symmetrical operation of undermining factors suggests an independent and explanatorily prior notion of responsibility. This notion requires that the outcome be voluntary and intentional and unmistaken. I think these are the necessary conditions on being responsible, and I take up the task of defending this claim in the next chapter.

[121] Though, on my view, it wouldn't be *incorrect* to blame Manny for cracking the safe, or Otis for burning the field. We just take there to be decisive reasons against doing so. More discussion of this feature of our blaming practices will occur in Chapters 3 and 4.

Chapter 4: The Negligence Objection

4.1. Introduction

AT THE END OF CHAPTER 3, I stated the necessary conditions on responsibility for outcomes. I claimed that to be responsible for an outcome, the outcome must (1) be brought about voluntarily, (2) brought about intentionally, and (3) brought about without mistake.[122] The Voluntariness Condition is met so long as the outcome is brought about by an action explainable by a belief-desire set. The Intentionality Condition is met so long as the outcome in question is believed by the agent to be a possible result of that action. And the No-Mistake Condition is met so long as all relevant beliefs about facts necessary for generating an evaluation by the relevant standards are true.

In this chapter, I defend the claim that all three conditions are necessary. In particular, I want to rebut a suggestive objection that claims that one can be responsible for an outcome even in the absence of one of my conditions. The *Negligence Objection*, as I'll call it, claims that agents who cause harm due to their negligence fail to meet one

[122] These are the abbreviated statements of the conditions that I will refer to most often.

of my conditions on responsibility, and yet they are nonetheless responsible for the harm. Thus, my conditions are not necessary after all.

The objection rests much of its claim on the fact that when someone is negligent and, say, injures someone, we hold the agent responsible for those injuries and tend to judge him blameworthy. Indeed, blaming people for their negligence, or thoughtlessness, or carelessness, is an all too common occurrence. It is all the more surprising, then, that very little has been said about responsibility for negligently produced harms outside of discussions in legal theory. In fact, the moral responsibility literature has been virtually silent on the matter.[123] Such a pervasive feature of our blaming practices calls out for an explanation, especially if it's to serve as support for an objection to a theory of responsibility.[124]

In this chapter, I'll outline what's distinctive about negligence, and how one can mount an objection to my account by examining negligent agents (Section 2). Negligence is characterized by the lack of certain conscious mental states, so if negligent agents are responsible for what they do while lacking these mental states, then no account of responsibility could require those mental states without modification. In this way, the Negligence Objection actually poses a general problem for *any* theory of responsibility. Next, I'll argue that the standard way of explaining responsibility in cases missing requisite mental elements, what has come to be known in the literature as "tracing," fails to explain responsibility in cases of negligence (Section 3). It follows that either we'll

[123] One especially notable exception is Zimmerman [1986]. But Zimmerman's treatment begins by creating a special instance of negligence, one in which the agent previously thought about the possibility of harm, and then arguing that we can explain responsibility in such cases. Given his limiter, he misses most cases of simple negligence, where the agent never before consciously considered the risk of harm such conduct might pose.

[124] Manuel Vargas calls similar attention to negligence in Vargas [2005].

need a different explanatory story, or we should reject the claim that negligent agents are responsible for the outcomes they bring about. Next, I'll give some considerations for rejecting the first disjunct (Section 4), and then propose a model for handling negligence intended to reduce worries associated with the latter option (Section 5). Finally, I'll compare cases of negligence to cases of simple inadvertence to lend some support to this alternative view of negligence (Section 6), and offer a brief conclusion (Section 7).

4.2. The Nature of Negligence

NEGLIGENCE CONSTITUTES a special class of cases. Unlike harms that agents bring about on purpose, or knowingly, or even recklessly, negligently produced harms are brought about because of an absence of care. To highlight what's special about negligence, I'll contrast it here with recklessness. When an agent acts recklessly he consciously disregards the risk of harm his conduct poses. Though he recognizes that his action will create a "substantial and unjustified risk of harm,"[125] he acts anyway. Should he in fact cause harm as a result, it seems as though he is responsible for doing so. For example, suppose Spencer is late for an important appointment, and so he is speeding down the highway, swerving in and out of traffic, trying to make up time. He realizes that this is dangerous behavior, driving in such a way greatly increases the risk of harm, but he continues nevertheless. Sure enough, he causes an accident, and injures another driver. Speeding Spencer is reckless; for he has considered the risk his conduct involves and *consciously* disregarded it in favor of acting anyhow. Reckless agents satisfy my Intentionality Condition, since they at least foresee the possible occurrence of the outcome in question.

[125] This is how the Model Penal Code characterizes recklessness, but it accords well with common usage.

Negligent agents, in contrast, risk harm by not taking sufficient care in acting. They unreasonably fail to pay to attention to the possible consequences of their conduct, and thus substantially increase the risk of harm such conduct poses. But they don't do this consciously. That is, they *fail* to exercise due care; they should be paying attention but don't. For example, suppose that Nate, tired from waking up early, is backing out of his driveway. His thoughts turn to his meetings that day, and his attention is partially focused on a radio commercial. Due to his inattention, Nate doesn't see a child walking to school and so hits him, breaking the child's leg. Nate is negligent: he fails to pay proper attention to what he is doing and thereby risks harm to others. It seems Negligent Nate[126] fails my Intentionality Condition, since he does not actually foresee the possible occurrence of the outcome in question when he acts. And if he's responsible, then he stands as a counterexample to my three conditions being necessary for responsibility.

The requirement of consciously disregarding the risk of harm is crucially important, for it serves as a dividing line between recklessness and negligence. Negligence, like recklessness, involves engaging in conduct that risks harm. But negligence no longer requires consciously entertaining the risk one's conduct poses. It only has to be the case that one's conduct is unreasonably risky, not that one acted in the recognition that it was so. Thus, negligence abandons the element of conscious consideration involved in reckless behavior (and intentional conduct, for that matter). To do x recklessly requires consciously entertaining that x might result from your directly intended action. But to do x negligently is to do x as a result of *not* consciously entertaining the risk of x given one's directly intended conduct and refraining from that

[126] Negligent Nate's name here is not meant to suggest a general character trait, but merely to help remind the reader of the specifics of the case when it is later brought up.

conduct. Thus, we should consider negligence as being importantly different from recklessness (and directly intended action) in that it does not require a conscious entertaining of the harm or the risk of harm. Indeed, it is characterized by the *failure* to consider the risk. The hallmark of negligence is the lack of a conscious element.[127]

But my account (and, indeed, most accounts) of responsibility require at least some conscious mental element tying the agent to the outcome in question. In my case, it would appear that Negligent Nate fails the Intentionality Condition, thus it follows that on my account Nate isn't responsible for the child's injuries. The Negligence Objection argues that since negligent agents are responsible for the harms they negligently produce, Nate is responsible for the child's injuries, and my account gets the wrong result. More than this, the Negligence Objection becomes a general objection to most extant accounts of responsibility,[128] since they require some conscious mental element be present, and negligence is characterized precisely by the lack of such an element. The Negligence Objection is thus not a unique objection to my particular account, but a general problem facing all theories of responsibility.

4.3. Tracing and Negligence

IF WE ARE TO BE responsible for the products of our negligence, as the Negligence Objection suggests, then it must be the case that responsibility doesn't always require a connection between harms and some conscious mental state. This is no special problem,

[127] It is worth noting that I'm using "negligence" semi-stipulatively. There is no doubt, I think, that we do blame people for harms they bring about due to their unconscious inattention. I'm calling such agents "negligent." Of course, there may be other linguistically legitimate uses of "negligence" than mine here; naturally, my arguments won't necessarily extend to such cases.

[128] This qualifier will be discharged below, where I again discuss Quality of Will approaches, which try to avoid reliance on conscious mental states in their explanations of responsibility.

one might think, because we already have a strategy for explaining responsibility in the absence of a conscious mental element. We call it *tracing*, because responsibility for some conduct without the conscious mental element can be "traced back" to some previous decision or action that does have the conscious mental element. For example, Sven is drinking at a bar, and has one (or maybe more) too many. Sven is sloshed. Nevertheless, he drives toward home. En route, and due to his drunkenness, he hits a family sedan, seriously injuring all four passengers. Suppose that Sven is sufficiently intoxicated that he lacks the relevant conscious mental states for responsibility at the time of the crash. If we want to maintain that Sven is still responsible for hitting the sedan, then his responsibility for that action must be located elsewhere. This is where tracing comes in. We can "trace" his responsibility for the crash to his responsibility for a prior act that contributed to the crash. Sloshed Sven elects to drink to excess, and as a result of this choice, he hits and injures a number of people. His choice left him drunk, and therefore severely impaired with respect to his ability to control his conduct, to recognize its relevant features. In this case we can say that the initial choice creates a condition of impairment that later clearly contributes to some harm. The above is, roughly, the structure of tracing.[129]

It is worth noting that this is a *general* way of explaining cases like Sven's. It is a strategy that can be adopted by all theories of responsibility to handle cases in which the usually requisite mental states are not present. We can preserve responsibility for the harm so long as we can trace the harm back to some prior action which did include the

[129] Tracing has been an explicit interest of theories at least since Dennett [1984], and especially in responses to that work in Ekstrom [2000], Fischer & Ravizza [1998], and Kane [1996]. The core notion is also discussed by van Inwagen [1994] and, in response, by Petit [2002] and Vander Laan [2001]. For more recent work on tracing and its prospects, see Fischer & Tognazzini [forthcoming], McKenna [2007], and Vargas [2005].

relevant conscious mental element. Then all that's required is that the agent satisfy the conditions on responsibility for *that* prior choice or action, and responsibility can be transmitted to the later outcome.

Tracing plainly will not work, however, in cases of negligence, for in such cases it is difficult to demonstrate what the initial choice is. Sloshed Sven's accident is largely due to his drunkenness, for which, by hypothesis, he was responsible.[130] Tracing claims that in such circumstances responsibility is preserved to (at least some of) the outcomes produced by his drunkenness, even if at the time of those later actions Sven lacks the conditions normally required for responsibility. This is because Sven, roughly, chooses to get drunk. Contrast this case with Negligent Nate. He doesn't choose to be inattentive, nor does he do anything for which he is responsible that also obviously creates the condition of his inattentiveness. First, recall that the inadvertence associated with negligence is a failure, and as such is characterized by the lack of conscious mental awareness of potential harm. It is not the disregarding of any consideration nor does it involve the realization of risk. That is how we distinguish it from recklessness. Negligence is crucially *defined* by the absence of consideration, not its conscious dismissal. Second, we need to consider the potential contributing factors to his inadvertence. In the example as stated, Nate is groggy, part of his attention is focused on his meetings that day, and he is partially distracted by the radio. These do not seem to resemble choices in the way that Sven's continued drinking is a choice. At the very least, one needs to show how these three factors would create a condition of inadvertence (as Sven's drinking creates his drunkenness), one characterized by its non-conscious nature.

[130] If one doubts his responsibility, we'll just have to come up with another case, since his will no longer illustrate how tracing is supposed to work.

Nate plausibly does not choose to think about his meetings that day, nor is it obvious that such thinking would lead one to be inattentive. Often thoughts simply occur to us, and it is certainly common to have thoughts about one's upcoming day just "rise to the surface" of conscious thought. It also doesn't seem as though Nate chooses to be groggy. He may choose to go to bed at a given time, and also to wake up, but we need not suppose that this entails he chooses to be groggy. Perhaps he got plenty of sleep and allotted plenty of time to "wake up" before leaving. It could nevertheless be the case that despite his precautions he is still groggy when he leaves the house. It also isn't clear that he chooses to drive while groggy; that is, he may not realize that he is groggy. And if he doesn't realize it, and he took the aforementioned reasonable steps to avoid grogginess, this would seem sufficient for undermining his responsibility for his grogginess. Lastly, though he may choose to turn on the radio, he doesn't choose to be distracted by it. Often background music or commercials can simply grab a hold of our attention, even when we wish it not to, and it can be a hard matter to predict when this will be the case.

Even should we think that Nate should have done more to guard against distraction, we could simply modify the case to remove the worries. Perhaps he is distracted by the sunrise, or a bird nearby, instead of the radio. Perhaps he isn't even groggy; he just gets distracted or otherwise fails to pay attention and so hits the child. The point is that simply failing to look behind him and hitting the child will be sufficient for demonstrating negligence, but nevertheless there will be no conscious choice to trace responsibility back to. Thus, it seems that tracing is an insufficient explanation for the agent's responsibility for negligently produced harm.

4.4. A Different Explanation for Negligent Responsibility

IF I'M RIGHT, THEN we can't appeal to tracing, the standard explanation for aberrant cases, to explain responsibility in instances of negligence. It follows that either negligence requires an exceptional explanation, or else we ought to reject the claim that negligent agents are responsible for the harms they bring about. In this section, I'll first argue generally that an exceptional explanation ought to be rejected (4.1). Then I'll rebut a specific suggestion that there is a general account of responsibility that can handle negligence cases in the same way as other cases (4.2). If such a view succeeded, then the Negligence Objection could be repurposed as an argument in favor of this general view.

4.4.1. Explaining Responsibility

MY ARGUMENT ADOPTS a simple constraint on explanations. If two cases share a common feature, then the explanation of that common feature for the two cases ought to be unified in some way. This can be applied to the issue before us here. All actions for which an agent is responsible share a common feature, namely, responsibility. If both Sloshed Sven and Negligent Nate are responsible for the harms they bring about, then we should expect the explanations of why they are responsible to be unified in an interesting way. Furthermore, given that one who performs an action on purpose, with a given outcome as his intended aim, is undoubtedly responsible, then the explanation for why this agent is responsible should be unified with the previous two cases. In short, if there's something it is to be responsible for things, then the only satisfactory explanation of this fact will be unified across cases of responsibility.

Explanation by tracing is an attempt to give such a unified account. Tracing tries to explain all cases where the relevant conscious states were absent by reference to those same conscious states in some prior action. Tracing may succeed in doing this between paradigmatic cases of responsible action and cases like Sloshed Sven's. Tracing bridges the gap between the two types of cases by showing where in Sloshed Sven's case the relevant conscious mental element is. Tracing purports to show how the cases are alike, and how *this* similarity explains responsibility in both cases. But tracing cannot bridge the gap between Sloshed Sven and Negligent Nate. Nate lacks the conscious mental element and there is nothing to trace back to. So the problem isn't just that tracing fails to explain responsibility in cases of negligence, it is that whatever the explanation is for negligence, if negligent agents are responsible, then the explanation should be unified with the explanation of responsibility both in cases like Sven's and in cases of paradigmatic responsible action. But this looks like a very unpromising project, for it seems that responsibility in paradigmatic cases is due to the presence of certain conscious mental states, just those sorts of states the absence of which characterize negligence.

4.4.2. Quality of Will Approaches

TRACING FAILS TO GIVE a unified explanation. But perhaps the Negligence Objection merely highlights a consideration in favor of an alternative general explanation of responsibility across cases. There is a family of views for explaining responsibility that gives priority not to the notion of control or conscious mental states, but rather to the quality of will an agent's conduct "manifests."[131] On this view, that an agent's conduct manifests an ill quality of will is sufficient for demonstrating responsibility. If Mad Max punches out the bartender, then this conduct manifests a quality of ill will on Max's part towards the bartender. The idea here is that conduct that isn't chosen can still plausibly manifest qualities of will. For instance,[132] a husband who routinely fails to consider his wife's interests and to at least occasionally place them above his own seems to express some quality of ill will towards her. In another example, repeatedly forgetting a close friend's birthday seems to manifest a lack of consideration towards him, and this ill quality of will seems sufficient for generating responsibility. When we do things that manifest ill qualities of will, so the view goes, we are the legitimate targets of certain kinds of criticism on the basis of that conduct (the kinds of criticism intimately associated with responsibility).[133]

One might think that such views have an easier time explaining negligence. It doesn't matter that Negligent Nate doesn't choose to run the child over, his failure to pay adequate attention nevertheless evinces a quality of ill will. Naturally, it's not as bad a quality of will as if he had harmed the child intentionally, or knowingly, say. Nevertheless, failure to pay attention when one is engaged in activities that pose a risk of serious harm displays a lack of consideration for those who you risk harm to. So,

[131] This is a major component of the Strawsonian Approach I argued against in Chapter 1. Here, I am considering the Negligence Objection as a possible response to my conclusions from that chapter. If negligent agents are responsible and Strawsonian accounts can explain this fact, then this claim would weaken my criticisms against the view there. The classic statement is Strawson [1962], who speaks in terms of actions "reflecting" and "expressing" qualities of will (p.63). For a more recent and developed presentation of such a view see Wallace [1994], who adopts the terminology used above. For a different application of a quality of will approach, applied to responsibility for attitudes, see Smith [2005].

[132] The following examples are from Smith [2005].

[133] Smith, A. [2005], p.243.

Negligent Nate displays ill will towards the child, and is thus responsible for harming him.

I think this approach fares no better than tracing in explaining responsibility in negligence cases. To see this, we first need an account of what it means for an action to "manifest" a quality of will. Unfortunately, the main proponents of these views say very little about what the "manifesting" relation is. As I see it, then, there are two main options: "manifest" could be a causal relation or an evidential relation. There is some evidence for either gloss. Some authors use "manifest" simultaneously with "express," which often looks causal, but the use of the relation on these views is often to draw inferences from actions to qualities of will. No matter our gloss, however, on neither understanding will quality of will approaches be able to explain negligence responsibility. Indeed, they both fail for the same reason.

My preferred reading of "manifest" is as an evidential reading because the role it plays in the theory is to support inferences from actions to qualities of ill will. Unfortunately, on this reading quality of will theories will fail to explain negligence responsibility. On the evidential reading, we are supposed to think that Negligent Nate's conduct evinces an ill quality of will. He should have paid more attention to what he was doing, and the fact that he didn't is evidence that he doesn't give the appropriate consideration to those who he risks harm to in operating a vehicle carelessly. But this conclusion is too strong. The power of the evidential relation surely rests on the reliability of the inference from conduct to ill qualities of will. The reliability of such an inference requires, it seems, some regularity in its connections. Notice that in sketching the examples above I used words like "repeatedly" and "routinely." Of course, any conduct can count as some evidence for the underlying quality of will, but we generally require more before we’re justified in actually drawing the inference. In order to justifiably draw the inference we need something like a pattern of response. But ascriptions of responsibility in cases of negligence do not rest on regularities. Negligent Nate could have at all times previously been the paragon of careful driving. This is counterevidence, it would seem, for thinking that in the particular case in question, Nate manifests ill will towards anyone, even the child. Nevertheless, one transgression is sufficient for negligence, and if negligence itself is to be sufficient for responsibility, then it seems that quality of will views fare no better in explaining it.

Quality of will views will fail on the causal reading for much the same reason. On this reading we are supposed to think that Nate’s negligence, his failure to pay attention, is caused by some ill quality of will, either towards the child or in general. It seems to me we should be initially skeptical of such a claim, for Nate doesn’t even know that the child, or anyone else, is there. Nor is he even consciously thinking about backing up, so we might seriously doubt that his current frame of mind is such that his attitudes towards others, whatever they are, would be engaged to cause his lack of attention. More importantly, however, Nate could very well hold quite positive attitudes toward the child he injures. Perhaps the child is a neighbor’s son, who mows Nate’s grass and shovels his sidewalk when it snows. Nate is very grateful for the son’s work, thinks him a fine young man, etc. If this were the case, it would be hard to claim that Nate holds any ill will toward the child, much less that that ill will caused his failure to pay attention. Indeed, it seems our causal judgments will be tied up again with the reliability of our inferences from the conduct to the agent's attitudes, and will therefore be subject to counter-evidence of the above sort.

No matter the gloss we give to "manifest," quality of will approaches fare no better at explaining responsibility for negligently produced harms. Quality of will accounts seem to require some regularity of conduct. Repeatedly disregarding someone's interests might count as sufficient evidence for an ill quality of will, but a one-time offense would be insufficient. This is especially true anytime we have counterevidence involving positive qualities of will. But once is enough for an ascription of responsibility in negligence cases. Even someone who in all other respects is the acme of consideration and care can be responsible for negligent conduct, it seems, should he fail to pay proper attention in just one instance. Quality of will accounts, therefore, cannot explain responsibility in cases of negligence.

4.5. Rejecting the Claim that Negligent Agents are Responsible for Harms

IF I'M RIGHT, THEN tracing fails as an explanation for responsibility for negligently produced outcomes. It fails because it requires tracing responsibility back to some conscious mental element, usually a choice or action. But negligence is defined by a *lack* of conscious mental states, by an unconscious inattention, and there need be no choices or decisions that contribute to that inattention. Thus, tracing is unable to help us explain responsibility for negligently produced outcomes. If I'm right, then, we are seemingly left with two options: either revise our accounts of responsibility in an effort to explain the outcomes of negligence,[134] or drop the intuitive conclusion that negligent agents are

[134] Manuel Vargas has discussed some different problems with tracing, and its relation to negligence, in Vargas [2005].

responsible for the harms they produce. I argued in the previous section that the first option is untenable.

The other option one can take is to reject the claim that Negligent Nate is responsible for the child's injuries. If that were the case, then we should expect tracing to fail as an explanation. Indeed, no satisfactory explanation could be given that would unify such cases with paradigmatic cases of responsible actions. The burden on such a move is showing why it seems that negligent agents are responsible for the harm they do. Why does the Negligence Objection's claim that Negligent Nate is responsible seem so plausible?

Part of the reason no doubt is that we do tend to hold negligent agents responsible for their conduct. Most people would blame Negligent Nate for injuring the child; after all, it was his fault. And to claim that he is at fault is to criticize Nate. But we can find him at fault, I submit, without requiring that he is responsible or blameworthy for the harm.

Even though he is at fault, it doesn't follow that he's responsible or blameworthy for the harm. To be at fault may well require less stringent standards than being responsible, in the sense necessary for blameworthiness, requires. This distinction is well-mirrored in a legal distinction between criminal liability and civil liability.[135] Criminal liability seems to closely resemble moral responsibility. If one is criminally liable for something, then they are "sanction-worthy," subject to punishment or deserving of what would otherwise be objectionable treatment. This connection is not preserved in the civil law. To be civilly liable is merely to be designated as the appropriate individual for compensating injured or harmed parties. The distinction is supported by the differing

[135] A similar point is discussed in "Notes" [1972], esp. pp. 976-979.

aims of the criminal law and of the civil law, respectively. They seek answers to different questions. The criminal law is concerned with determining who should be punished. Since punishment often involves harms, it is the sort of treatment that requires justification. Here, the criminal law surely shares something in common with our practices of blaming, which also seem to call for special justification. Indeed, it seems proper to view the criminal law as a distinctive instantiation of our blaming practices. The special justification punishment requires involves a high standard for connecting the individual to the offense, something akin to moral responsibility.

The civil law, on the other hand, is concerned with determining who should pay. Civil trials (here I confine myself to torts) begin with the acknowledgement that the plaintiff has been injured or otherwise harmed, and his burden is to demonstrate that the defendant caused the harm by acting inappropriately.[136] Here, the civil law does not need as stringent a justification for imposing the burdens of compensation on the defendant, as it would if it were punishing him. But civil verdicts do not lead to punishments; they lead to penalties, usually financial, that serve to compensate the plaintiff for damages suffered. We need not require that a defendant be morally responsible for some harm in order to be the individual judged best placed to compensate the one harmed. We only need to show that the defendant caused the harm and has no defense for why he shouldn't pay for it. The presumption in civil cases, then, is that should the plaintiff show that the defendant acted inappropriately (e.g., unreasonably) and thereby caused the harm, this is sufficient for sticking the defendant with the bill.[137]

[136] This is a gross oversimplification, but the details do not matter for my point.

[137] Indeed, sometimes showing cause isn't even necessary. A number of civil cases have been decided by an appeal to fairness: that it would be unfair to impose the burdens on the injured party, and instead the defendant ought to pay, even though the plaintiff cannot prove that the defendant even caused the harm,

If negligent agents are not really responsible for their conduct, then, perhaps we can retain and justify many of our attitudes and practices towards such agents by appealing to a less stringent connection, something more like civil liability. It seems as though moral responsibility requires a conscious mental element, but negligence lacks any relevant conscious mental state. But negligence also involves failing to appreciate the risk of one's conduct, a failure that seems in many cases to be unreasonable. We can rightly demand of Nate that he look behind him, and when he fails to, he must acknowledge his failure as being inappropriate, as violating our standards of care. He doesn't do what he's supposed to do when driving. Now, if my argument has been successful, it would be a mistake to blame Nate for injuring the child, as he is not responsible for it. However, we can acknowledge that he is at fault, that his conduct was nonetheless criticizable, and that he is required to compensate the child for his injuries. Negligent Nate does something "wrong;" this is part of what it means to be negligent. He fails to pay sufficient attention and acts unreasonably as a result. But we can acknowledge his fault without an ascription of responsibility for the harm.

Allow me to elaborate briefly on this last point. When we admonish Nate, we can do so merely by referencing his fault with respect to his failure to pay attention, for we are simply highlighting the fact that he has failed our expectations to avoid causing harm. We require individuals to look when they back out of driveways (in part because we expect them to recognize the dangers posed by operating vehicles). When individuals fail this requirement, we can rightly criticize that failure. We can even point to the child's injury as a relevant component of that criticism. But in doing so, I think, we are merely

much less that he (or they) is responsible in any "deep" sense. See, e.g., *Summers v. Tice*, and *Sindell v. Abbott Laboratories*.

reemphasizing why we have such requirements in the first place. You're expected to look behind you when you back out of a driveway so as not to cause injury to others. And when one fails to pay sufficient attention and thus causes injury, this fact is a valid source of criticism. But it need not involve any claims about responsibility for the harm or condemnation of the individual.

It may of course be the case, as I noted above, that consistent failures of these requirements to exercise care do point to certain attitudes and values that *are* proper objects of responsibility.[138] Still, in isolated cases of negligence, we need not infer anything about, say, Nate's attitudes towards the child, nor think that his lapse of attention evinces a lack of concern or an ill quality of will. Indeed, Nate may be despondent over the harm he's caused, and he may express regret. There is nothing wrong with such attitudes. We can regret not taking a chance that would have paid off big just as easily as regretting not hearing a cry for help because we were guiltlessly listening to our iPod. My only caution is that either regret does not depend on being responsible for the object of regret, or that regret is technically misplaced in the latter case, just as it is in instances of negligence. But there may be good reason to encourage a certain amount of negative attitudes towards instances of negligence, if only to reinforce the importance of our shared standards of care. To the extent that we can emphasize dangerous lapses in attention as things to be avoided, so long as we can reaffirm the importance of taking due care in everything we do, we ought to do so. Negligence often risks quite serious harm, which is to be avoided, and our criticism of negligent agents can be justified by appeals to education or simply reaffirming a commitment to the avoidance

[138] This thought is along the lines of the view defended by Smith [2005].

of harm. But we shouldn't confuse such a justification with an ascription of responsibility for the products of negligent conduct.

Treating negligence in this way, as modeled on the civil law, preserves much of what I take is central to our reactions toward negligent agents. Isolated transgressions are the proper object of admonishment, largely in an effort to educate or to deter future transgressions or even to reaffirm a shared commitment to our standards of care. Naturally, chronic transgressions may point to a defective character trait, or some other proper object of moral condemnation. But in the more typical instances of local negligence, we can retain much of our disapproval and negative reactions by appeal to the negligent agent's "fault," and give up a more stringent requirement of responsibility.

4.6. Support for the Alternative View

IN SUPPORT OF TREATING cases of negligence on the proposed model of civil liability, I want to compare cases of negligence with a structurally similar set of cases, namely, simple inadvertence. In cases of inadvertence, we don't typically ascribe responsibility for outcomes; indeed, inadvertence is a factor that seems to undermine responsibility. Explaining these cases together, then, pressures us to treat negligence like inadvertence, and therefore reject the claim that negligent agents are responsible for the harms they produced. Now it is likely the case that intuitions will still diverge about negligence cases and cases of inadvertence. That is, many will still judge negligent agents responsible, but not their inadvertent counterparts. We can explain this fact, however, by reference to what distinguishes negligent agents from their inadvertent counterparts. It is a difference which I think we'll show why we tend to treat negligent agents more harshly, but it is not a difference that will actually demonstrate them to be responsible.

4.6.1. Negligence as a Species of Inadvertence

WE KNOW WHAT Negligent Nate did, but let's compare him with a case from Chapter 3. Leadfoot Lenny[139] is at a party where a group of friends are gathered watching a movie. There are more people than seats, and some have gotten comfortable lying on the floor. Lenny gets up to get a soda from the fridge, and in the course of stepping around and over people he inadvertently steps on his friend's hand. He didn't mean to step on his friend's hand, but he was distracted by the movie, and so he did. Nevertheless, it seems Lenny's responsibility in this case is undermined due to his inadvertence, which makes his stepping on the hand unintentional.

We seem to treat negligence as preserving responsibility, whereas inadvertence seems to undermine it. But if this is the case, then there must be a way of distinguishing between Nate's case and Lenny's in a way that explains why Nate is responsible and Lenny is not. Otherwise, we're forced to conclude that inadvertence doesn't undermine responsibility, and this is a conclusion I should think we want to reject.

An initial observation about cases of negligence is that they involve a failure to do something the agent should have done; negligent agents should have paid more attention. We say, "Nate should have looked where he was going." Unfortunately, such claims are ambiguous. There are two possible interpretations for how to treat the "should have" clause in each, and each interpretation seems to fail to distinguish properly between

[139] Lenny's name here, like Nate's, is not intended to reveal a character trait, but only make clear the facts of the case when brought up later.

Negligent Nate and Leadfoot Lenny. On the first interpretation, the "should have" means simply that it would have been better had Nate done what he failed to do. It would have been better had Nate looked because then the child wouldn't have been hit. The problem with this first interpretation is that, while true of Nate's case, it is equally true of Lenny's case. It would have been better if Lenny had been more careful, because then his friend wouldn't have gotten his hand stepped on. So the first interpretation fails to distinguish between the cases at all.

On the second interpretation, the "should have" refers to some sort of standard that was violated. We require individuals to look when they back out of driveways (in part because we expect them to recognize the dangers posed by operating vehicles). Moreover, this seems to be an instance of a general duty to take extra care when engaging in activities that pose a risk of harm (or, perhaps, just a standing duty to take care in our conduct). The problem on this second interpretation is that we can always ask why non-responsible counterparts aren't under a similar sort of standing duty. Navigating around people lying on the floor poses a risk of harm. If a standing duty is sufficient for securing responsibility, then Lenny would be responsible as well. Granted, in this case, Lenny causes less harm than Nate, and is involved in an activity that poses a risk of less serious harm, but that shouldn't count against Nate. While our expectations are surely stronger in Nate's case, this seems insufficient for setting negligence cases apart as a distinct class of cases with respect to responsibility. After all, a standing duty governing a given activity applies to *all* those who engage in the relevant activity. But the relevant requirement here isn't driving carefully. If it were, it wouldn't apply to negligent agents whose negligence has nothing to do with driving, like a bricklayer who tosses defective bricks off a rooftop without looking.[140] But the bricklayer seems negligent in the same way as Nate is: he should have paid more attention to the harm he risked. And even if we have scores of standing duties finely individuated by activity, so that there's one governing driving, and another governing bricklaying, and another governing firing a gun, etc., this is only because they fall out of a quite *general* standing duty to take care in everything we do.

It might be thought that satisfying the standard may require different thresholds of care depending on the specific activity. So, when one is driving or bricklaying on a roof, one needs to take special care, but not when one is walking around prone people on the way to getting a soda. But I'm not convinced that this is really a demand for *more* care than in Lenny's hand-stepping case, and not just a different side of the same point about driving and bricklaying on a roof having the potential for more harm than stepping around prone people.[141] Nevertheless, even if we think you do have to be more careful in certain situations, this wouldn't distinguish between the cases on grounds that evince a difference in responsibility. For the mere violation of a standing duty of care wouldn't suffice to show an agent is responsible for the effects of such a violation. And if it did, Leadfoot Lenny would be responsible too.

[140] This example is from Hart [1968], and is often reused in discussions of negligence in legal theory. See also Zimmerman [1986].

[141] Bénédicte Veillet has pointed out to me that we license people for driving (Nate), whereas we don't license people for stepping around prone people (Lenny). This is true, no doubt in large part because of the importance we place on driving. It is obvious that such activities pose the risk of especially serious harm, more so than stepping around prone people. But I don't think this amounts to a claim about demanding more care in such instances, but rather a certification that those we license can demonstrate those skills that constitute taking care in the circumstances germane to the activity. Plus, the costs of requiring and regulating licenses are worth it, given the seriousness of the potential harm. This isn't the case in the matter of prone-people-stepping. I say more about these points below.

4.6.2. Negligence and Negative Expected Value

BUT WHY DOES THE intuition that Negligent Nate is responsible seem so strong? One important difference between Nate's case and Lenny's is that Nate causes much more harm. A broken leg is much worse than a trod upon hand (we're assuming the hand isn't broken). Additionally, there's only so much harm you can cause by inattentively walking around prone people, whereas the risk of serious injury by inattentively operating a car is much greater. It doesn't seem to me that considerations such as these can support differing judgments of responsibility, but they may help explain our differing reactions to the cases.

I don't think the severity of the harm caused can make a difference in responsibility. Suppose you knew that an agent caused some harm entirely by mistake, by making an entirely reasonable choice, say. If you were convinced that the agent wasn't responsible, would learning that it was a serious harm make a difference? To put the point generally, suppose we take a case in which it's obvious an agent isn't responsible and then ratchet up the seriousness of the harm brought about. Is there a threshold in which responsibility "appears"? I find the very suggestion to be implausible.

One reason to think it's implausible surfaces if we compare paradigm cases of responsibility. So, suppose Deirdre hates Emma and wants her to suffer. So Deidre waits for Emma to get off work, sneaks up behind her in the parking lot and beats her with a baseball bat. Emma suffers multiple contusions, a cracked skull, and a concussion. Deidre is responsible for Emma's injuries if anyone ever is. Now suppose that Fran hates Ginny and wants her to suffer. So Fran waits for Ginny to get off work, sneaks up behind her in the parking lots and pulls down her pants, causing Ginny to trip and fall. Ginny skins her knees and is humiliated in front of a couple coworkers who observe the attack.[142] Fran, too, seems responsible for Ginny's injuries if anyone ever is. Indeed, in comparison, Deidre and Fran seem equally responsible for their respective harms. The paradigm of responsibility is cases of intentional action aimed at a particular outcome as the intended end, where the end in fact occurs as intended. Deidre and Fran's actions fit the bill. Both are out solely to harm (in different ways) their targets, and this is precisely what they do. In fact, I don't see how one could be more responsible for an outcome then they are in their respective scenarios.

It should be apparent, however, that the harm done to Emma is much worse than the harm done to Ginny. Indeed, Deidre is certainly more blameworthy than Fran for what she does precisely because the harm Deidre brings about is so much worse.[143] But the amount of harm brought about isn't pertinent to the ladies' responsibility for bringing it about. Both are equally responsible in their respective scenarios. So, while the amount of harm brought about is certainly relevant for how blameworthy an agent is, it isn't relevant to her responsibility for bringing it about. Responsibility for outcomes is interested only with the relation between an agent's mental states (a psychological set explaining action) and a given outcome. That's why Deidre and Fran are both equally responsible for their respective harms, even while Deidre is more blameworthy for her harm than Fran is for hers.

This result can be applied to Nate's situation. First, let's change the case slightly, and let's call the new Nate, Nate*. Suppose everything is as before, but instead of not

[142] We can leave any evaluations of these coworkers aside.

[143] There maybe other relevant factors in play here. We may judge Deidre more blameworthy overall because we also think it's worse to intend to physically assault someone than to harass them a bit. But I want to limit the discussion above to comparative judgments regarding blameworthiness for what they do, blameworthiness for particular mental states notwithstanding.

seeing the child, he doesn’t see a child’s bike that’s been left on the sidewalk. So Nate* runs over a child’s bike. This is a much less serious harm than what Nate brings about. Is Nate* less responsible for running over the child’s bike? I don’t think so. Responsibility ascriptions seem importantly insulated from the amount of harm caused, as Deidre and Fran showed. So I think the answer one gives should match whatever one thinks about Nate’s case, and the same goes if Nate* runs over a prized rosebush or three hundred nuns. Since we should treat these cases alike, and I think Nate* isn’t responsible, Nate isn’t responsible either.

But this isn’t to let Nate off the hook in any of these cases. We would reproach him more strongly than Leadfoot Lenny, and many people would still react to him as if he were blameworthy for the harm. I agree that we tend to react this way. Moreover, I think we’re justified to act this way. We have all sorts of good reasons to admonish Nate. We have standing duties of care for a reason. When people fulfill the duty they tend to cause less harm. And we have very strong reasons to be in the business of seeking to limit harm in general. So I think we are likely justified in censuring Nate, reprimanding him for his failure, and chastising his causing of the injury (which wouldn’t have occurred without his failure), but I don’t think he’s actually worthy of blame for the harm. Our reactions to him for the harm may be justified because they reinforce the importance of the standing duty of care. It’s the equivalent of us saying, “This is why we have such a standard, to avoid harm like this! That’s why we demand you follow it!” But this response merely reinforces the fact that Nate is sensitive to our shared practices of demands and expectations, that he can modify his behavior and is a reflective agent, and it helps justify the standing duty itself.[144]

Consider an analogous aesthetic example. Suppose Julius is a recording engineer, recording the latest album for the heavy metal band Scooby Doom. It is Julius's job to manage the various inputs and outputs, the recording levels, etc. Julius neglects to check a vocal amp that was used for the last song but isn't in use on the song he's currently recording. As a result, there is a persistent hiss on the track.[145] The band will yell at Julius and complain that his neglect led to a poor recording of the song. They may even blame him. No doubt, their reactions will center on his neglect, on the fact that he failed to do something he was supposed to do. And in reacting in this fashion, they will no doubt draw heavily on the fact that the neglect caused a hiss, and the hiss ruined the take. But these facts do not establish Julius' blameworthiness for the hiss, so much as figure in explanations for why Julius' failure matters. It matters because such failures tend to lead to bad things, like the hiss on the track. And hisses on tracks are aesthetically bad outcomes. But Julius need not be responsible for the hiss in order for him to be criticizable for his failure, nor in order for Scooby Doom's admonishment of him, even for the hiss, to be justified.

Now we have an explanation for our differing reactions to Negligent Nate and Leadfoot Lenny, respectively. We tend to treat Nate more harshly because his failure of the standing duty of care is worse than Lenny's. We can understand this comparison in terms of the amount of harm risked by the activity each is engaged in. I am assuming

[144] And in some cases where an individual isn't sensitive to our shared practices, like a child perhaps, admonishment can serve to educate her about these practices.

[145] This example requires some artificiality, since they could always rerecord the track. If this is a bothersome detail, the reader is invited to create a similar example featuring the sound technician's inadvertence during a live concert leading to an aesthetically disastrous feature of the performance.

here that the probability of harm is equivalent in these two cases. The reader is invited to tweak the example suitably if the reader doubts that the probabilities are in fact the same.[146] This assumption will be important below. Now, Nate is operating a vehicle, which poses the risk for quite serious harm. If done without care, driving a vehicle can kill multiple people. And this is a quite serious harm. Lenny's activity, while it does risk harm, risks a much more minor amount of harm. One could dislocate a finger or perhaps even break a bone by having one's hand stepped on, but these are fairly minor harms. This comparison helps explain what the line separating Nate's negligence from Lenny's simple inadvertence is doing. It is denoting a difference in 'admonishability' for failing the standard by reference to the expected negative value of the harms posed by engaging in such activities without due care. In both cases, I think, the agent is reproachable for failing the standing duty of care. Both Nate and Lenny are criticizable in this respect. Indeed, as experience will reveal, we would admonish both Lenny and Nate for their failures. "Ouch, watch where you're going!" we might snap at Lenny. And in keeping with my story for our reactions to Nate, this response also has as its source the standing duty of care, and our providing it serves to reinforce that standard. But driving vehicles poses a much more serious risk of harm, because the sorts of harms that could result if one doesn't satisfy the standard are so much worse. Thus, Nate's violation is worse; it more flagrantly flouts the duty to avoid harm by engaging in a more dangerous activity without being sufficiently careful.

However, this isn't the only way to be more admonishable for failing the standing duty of care. The line between Nate and Lenny is drawn by the expected negative value

[146] We could do this artificially by example, but the examples would seem a bit contrived. I prefer to assume the equivalence of probability and work with more ordinary examples.

of the harm risked without due care. But there are two ways one can increase the expected negative value of the harm. The first way, as demonstrated in the case between Nate and Lenny, is to increase the amount of harm risked by the activity. Driving vehicles risks more harm than stepping around prone people. The other way to increase the expected negative value of the harm is to increase the risk of harm; that is, to engage in riskier activities. For example, shooting one's firearm without care in the middle of Montana risks a certain amount of harm (e.g., a gunshot wound that could lead to death). Shooting one's firearm without care in the middle of Manhattan risks the same amount of harm, but it poses a much higher risk of that harm actually occurring. The chances that you will shoot someone in the middle of Montana are quite slim when compared to shooting someone in the middle of Manhattan. So, failing the standard in Manhattan, in this instance, is worse than failing it in Montana, and so is a more admonishable offense.

4.7. Conclusion

THIS VERDICT MATCHES our reactions to these cases, I think, and captures what is essential in distinguishing negligence from inadvertence. Negligence involves the failure to pay appropriate attention when the expected negative value of the risked harms meets or exceeds some threshold. Inadvertence involves the same sort of inattentiveness, but the negative expected value of the harm risked is much lower. When one's carelessness increases the negative expected value of the harm risked by too much, that agent is negligent.

Moreover, my alternative model for negligence also gives us a unified explanation. It unifies negligence and inadvertence by explaining them in terms of a shared inattentiveness and the expected negative value of the potential consequences of the conduct. Of course, this means the alternative view rejects the claim that negligent agents are responsible for the harms they bring about. But given the fact that negligence is defined by its absence of conscious mental states, and responsibility's traditional dependence on conscious mental states, this seems a virtue of my proposal. Furthermore, the alternative model captures a natural explanation of our reactions to cases of negligence, and helps show how we are justified in adopting admonishing attitudes towards negligent agents. Given tracing's failure to explain negligence, and the general worry that any unified explanation could be given for negligence, I'm forced to conclude that negligent agents are not responsible for the harms they bring about, but are properly admonishable according to our shared standards of care.

So I don't think we lose much in rejecting the claim that Negligent Nate is responsible. We can still claim that he is worse than Lenny, in the sense of warranting greater admonishment, because his violation of the standing duty was more egregious. We are forced to admit that Lenny, too, is reproachable for his failure to be careful. After all, he's under the same standing duty as Nate. But this isn't really a cost of the view; indeed, it actually fits with our general experience of similar cases. Lenny fails just as Nate does, but since the expected negative value of the harm of his activity without due care (stepping around prone people) is relatively low, his carelessness in this case doesn't arouse our intuitions as strongly as it does in Nate's case.

I don’t suppose to have convinced everyone here. We tend to have strong reactions to cases of negligence, and I don’t have knockdown arguments against such intuitions. But as we’ve seen, a proponent of the Negligence Objection will have difficulty applying a unified explanation to cases of negligence, and it seems generally difficult to distinguish between cases of negligence and inadvertence in terms of their responsibility for the harms. On my model, however, I can show how the difference between Nate and Lenny, while not concerning responsibility, makes sense of our reactions in a way supported by the phenomena.

If I’m right, then inadvertence (i.e., unconscious inattention) does uniformly undermine responsibility for outcomes. More generally, if I’m right, this strengthens my claim that my three conditions, Voluntariness, Intentionality and No-mistake, really are necessary conditions on responsibility. I also think that cases of directly intended action show that these three conditions are by themselves sufficient. I defend this claim in the following chapter.

Chapter 5: Four Cases for a Missing Condition

5.1. Introduction

I DEFENDED THE NECESSITY of my conditions on responsibility in Chapter 4. In this chapter, I defend the view that they are jointly sufficient. My strategy here is to consider four separate conditions, any one of which it could be argued is necessary for responsibility. Lacking a necessary condition would mean that my conditions couldn't be jointly sufficient. My aim, therefore, is to show that each candidate condition is not necessary for responsibility. Thus, if I succeed, in order to ascribe responsibility nothing else need be shown other than that the agent in question meets my three conditions: 1) the outcome was brought about voluntarily; 2) the outcome was brought about intentionally; and, 3) the outcome was brought about without mistake.[147] I again run the risk of leaving something out.[148] But I think I tackle four conditions most often proffered as necessary (beyond intentionality and voluntariness, at least). I'm confident that if I have left

[147] As before, these are the abbreviated versions of these conditions. Full discussion of them can be found in Chapter 3, Section 3.4.

[148] This was a risk in Chapter 3, where I tackled what I took to be the most paradigmatic (and oft-cited) examples of considerations that undermine responsibility.

something out, it is not a condition of central importance, and thus I assume it could also be dealt with. At the very least, rejecting these four conditions suffices to defend my view against what I see as its four toughest candidate conditions.

What are the four candidate conditions? The first is a condition of normative competence. Roughly, some argue that if an agent doesn't know that what he's doing is morally wrong, he isn't blameworthy for it because he isn't responsible for it. The second condition results from recent arguments concerning manipulation of various forms. The result is a historical condition. These arguments purport to show that an agent's history leading up to an action matters for responsibility. Roughly, an agent's history must be free from certain manipulations or interferences in order for her to be responsible for the outcome. The third condition figures usually as a premise in incompatibilist arguments. The claim is that since what one does at a particular time depends on how one is mentally at that time, to be responsible for what one does requires being responsible for how one is. Briefly stated, according to this view responsibility for outcomes requires responsibility for one's mental states and traits. I will refer to this as the ultimacy condition, as it is usually discussed in terms of ultimate responsibility for the things agents do. Finally, the last condition is that an agent must have genuine alternate possibilities to be responsible. It must be the case that the agent could have done otherwise. This is perhaps the most (in)famous and controversial condition cited as necessary for responsibility.

This chapter proceeds quite simply. In the next four sections I introduce a proposed candidate condition, and then argue against it. I'll proceed as indicated, arguing against normative competence (Section 2), a historical requirement (Section 3), the ultimacy condition (Section 4), and the alternate possibilities condition (Section 5), in succession.

5.2. Normative Competence

SOME HAVE GIVEN an argument that for one to be responsible for some outcome, one must know that what one is bringing about is morally good or morally bad.[149] We can extend the argument, I think, to include additional evaluative competence, such as artistic competence. The claim then becomes that one must know the normative evaluation of one's outcome to be responsible for it. One must know that it is morally bad, or artistically good, or scholastically bad, etc., in order to be responsible for it. In short, responsibility requires normative competence, knowledge of the evaluative status of the outcome in question according to the appropriate normative standards.[150]

I begin with a rough sketch of an argument, suggested by Gideon Rosen's view, augmented to speak more directly to my own account (2.1). Next, I present my rebuttal to this argument, borrowing on Pete Graham's own response to a similar argument (2.2). Finally, I'll conclude with some observations about what the argument does accomplish (2.3).

[149] For example, see Rosen [2002].

[150] One might think that normative competence is too broad. For example, there are norms of rationality. I think rationality norms are likely excluded from normative competence arguments. At least, including them would weaken any argument for a normative competence requirement, since mistakes of irrationality don't seem to be the sort that would upset responsibility. I think we do better, and stay truer to the intended scope of such arguments, to construe normative competence in a narrower way, as pertaining to norms of value, such as moral or artistic value.

5.2.1 The Argument From Moral Ignorance (AFMI)

THE ARGUMENT, IN BRIEF, suggests that just as ignorance of the facts usually excuses one from blameworthiness, so too can ignorance of the moral quality of one's act. In short, one is not blameworthy for bringing about x provided one did not know that he ought not to have brought about x.[151] This argument has some intuitive weight. Most of us think there are moral facts of the matter. What one ought to do in a given situation is a moral fact. So why limit instances of ignorance of the facts of the case to only the non-moral ones? The argument purports to show, by illustration, that there are at least some cases in which ignorance of what the agent ought to do excuses the agent from blameworthiness. If that's the case, then it seems that responsibility requires moral competence – at least a belief that one ought not be doing what one is doing.[152,153] If the argument succeeds, then, at the very least, my No-Mistake Condition requires revision to include normative facts along with the non-normative ones. I turn now to the argument.

Let's begin, as usual, with some cases. Consider the following two cases:[154]

[151] Rosen, for example, limits his discussion to instances in which ignorance of what the agent ought to do must be itself non-culpable in order to excuse. I argued against such restrictions in Chapter 3, but I'll retain his caveat for the purposes of my argument here. So nothing I argue for here should turn on my conclusions from Chapter 3.

[152] Obviously, the 'ought' in play here is a moral ought, since the argument doesn't aim to show that blameworthiness requires acting against what one believes one 'ought' to do in any respect. One may know that it is immoral to cheat but believe that he ought to, nonetheless, since it will increase his fame, say.

[153] It isn't clear whether the argument seeks to show a requirement of moral *knowledge* or not. That is, whether one must have something like a justified true belief about what one ought to do, and act contrary to it. Such a reading would also seem to require, given my comments above, that the moral domain fixes a fact of the matter about what one ought (all things considered) to do. I leave aside these various worries, and operate on the assumption that, at minimum, the Argument From Moral Ignorance seeks to show that ignorance of the moral nature of what one is doing excuses, and suggests some revision of my No-Mistake Condition.

[154] These are taken from Rosen [2002].

Slave Owner: Consider an ordinary Hittite lord. He buys and sells human beings, forces labor without compensation, and separates families to suit his purposes. Needless to say, what he does is wrong. The landlord is not entitled to do these things. But of course he thinks he is. Moreover, we may imagine that if he had thought otherwise, he would have acted differently. In that case he acts from moral ignorance in our sense. That much seems clear. It also seems clear that his ignorance is not straightforwardly grounded in factual ignorance. Unlike race slavery in the Americas, ancient Near Eastern slavery was not supported by myths about the biological or psychological inferiority of the slave. One became a slave through bad luck or imprudence; in principle the status could befall almost anyone. It is less clear to what extent this ignorance was grounded in false religion. The evidence suggests, however, that there was no perceived need for theological rationalization. The institution of chattel slavery was *simply taken for granted*. Questions about its administration were generally conceived as questions of civil law to be settled by convention or royal edict without recourse to higher principles.[155]

And,

[155] Rosen [2002], pp.64-65.

Sexist: Smith is a run-of-the mill American sexist circa (say) 1952. Like any decent middle class father he has encouraged his sons to go on to college, setting aside money for the purpose. But like any run-of-the-mill sexist he has done nothing comparable for his daughters. This differential treatment is not malicious. But it is unfair and therefore wrong. But of course Smith doesn't know this. He doesn't know that his daughters deserve equal consideration in this respect. …. [Moreover,] [l]et's suppose that if you had asked Smith at the time why he was treating his daughters differently, he would have said, 'Because they're girls,' as if the sufficiency of the answer were self-evident. Smith is the sort of complacent sexist who takes it for granted that his sons have legitimate expectations to which his daughters are not entitled (and perhaps vice versa). Let's suppose in addition that this commitment is not based on some sort of theory—some bit of bad religion or bad science. Let's assume, in other words—and this is hardly unrealistic—that Smith believes what he believes because he finds it obvious, and that he finds it obvious because he was raised to find it obvious and because the people he takes seriously find it obvious. The idea that gender matters in this way thus functions for him as an undefended axiom of moral common sense.[156]

[156] Rosen [2202], pp.66-67.

Proponents of the Argument from Moral Ignorance claim that neither the slave owner nor the sexist are blameworthy for what they do. The Hittite could not be expected to know that his treatment of his slaves was wrong. As Rosen puts it, "Given the intellectual and cultural resources available to a second millennium Hittite lord, it would have taken a moral genius to see through to the wrongness of chattel slavery."[157] It would be unreasonable of us to hold him responsible. So while we might decry a universe that would allow such an injustice to take place, "it makes no sense to hold this injustice against the [slave owner] when it would have taken a miracle of moral vision for him to have seen the moral case for acting differently."[158] Similarly, Rosen thinks that it would be unreasonable for us to blame the sexist for doing "what seemed reasonable given everything [he could] plausibly be expected to have known at the time."[159]

The defense here seems to be that we would be unreasonable if we blamed those who acted out of a reasonable moral ignorance.[160] At other times, Rosen suggests that it would be a "mistake" to blame those like the slave owner and sexist.[161] Given these facts, it would be inappropriate to blame the agent in question. If it would be inappropriate to blame the agent, then the agent isn't responsible.

[157] Rosen [2002], p.66.

[158] Rosen [2002], p.68.

[159] Rosen [2002], p.68.

[160] It is worth stressing here that Rosen, for example, goes to great pains to show that both the Hittite and sexist acted reasonably given their times. They were both as reflective as the common person of their day, and they didn't engage in any reflective negligence or the like. I assume with Rosen that these facts are established.

[161] For examples, see Rosen [2002], p.66,68.

5.2.2. Blaming as Something You Do and the Moralistic Fallacy (Again)

AFMI IS AN ARGUMENT by illustration. We are to consider the slave owner and the sexist, conclude that it would be inappropriate to blame either of them because they lacked important normative knowledge (or had specific false beliefs), and then reach the general conclusion that a knowledge condition on responsibility must include (at least) some moral knowledge (or correct beliefs).

Let's examine Rosen's defense of the claim that it would be inappropriate to blame the slave owner and the sexist. At bottom, Rosen follows Wallace in thinking that the main norms that govern the appropriateness of blame are norms of fairness. As he puts it,

> "It is unfair to blame someone for doing something if he blamelessly believes that there is no compelling moral reason not to do it. This principle is in turn supported by two more basic principles. It is unreasonable to expect people not to do what they blamelessly believe they are entitled to do, and it is unreasonable to subject people to sanctions when it would be unreasonable to expect them to have acted differently".[162]

So it would seem that blame is unfair when our expectations of their acting well are unreasonable. Thus, on Rosen's view, unreasonableness grounds unfairness.

As should be apparent, I don't think fairness is the right way to think about the appropriateness of blame. I argued as much in Chapter 2.[163] In brief, I think that what make blame inappropriate are the same things as what make praise inappropriate, namely, the undermining factors. I won't rehash that argument here. Instead, let's focus on the

[162] Rosen [2002], pp.74-75. I think it worth noting that it may just be false that we cannot reasonably sanction individuals for reasonable mistakes. It certainly isn't clear that no negative treatment of an individual for a reasonable mistake is warranted. I won't rely on this being the case, but it's important to note that Rosen's claim here isn't at all obvious. My thanks to Manuel Vargas for raising this point.

[163] See Chapter 1, Section 3.

two glaring mistakes I think Rosen's making here; mistakes encouraged by his adopting a practice-based view (like Wallace). First, notice that he confuses liability to blame with the unfairness of actually blaming. This is a crucial misstep. Second, this mistake leads him to conclude that considerations that would make blame unfair render one not liable to blame. I'll elaborate on these two mistakes in turn.

Liability to a response need not have the same conditions as the appropriateness of that response. Rosen, of course, understands liability to blame in terms of the appropriateness of actual blaming, but this is exactly what gets him into trouble. He succumbs to a version of the Moralistic Fallacy, as raised by D'Arms and Jacobsen and discussed in Chapter 2.[164] In their work, D'Arms and Jacobsen show that responses can "fit" their object (what I'm here calling "liability"), even when it would, in some sense, be inappropriate to actually engage in the given response. For a common example, something can be funny, even if it would be, say, morally wrong to laugh at it (perhaps because it's offensive to a minority). The distinction here is easily captured, I think, in terms of distinguishing between being worthy of a response and it's being the case that a particular response ought to be given.[165] Rosen confuses an agent's blameworthiness with it's being the case that one ought to blame him. It should be evident that the conditions that support the first clause can be different from the second. To blame someone is an action one engages in, and it is thus subject to a litany of reasons for action, even ones quite orthogonal to issues of responsibility. Blameworthiness, on the other hand, is a property or condition of an agent. For instance, if I will get a million dollars if I blame you for sneezing, then perhaps I ought to blame you, although there can

[164] See Chapter 1, Section 3.1.

[165] Recall the distinction found in both Chapter 2 (see also n.77), between 'blameworthiness' and 'it being the case that one ought to be blamed'. The same distinction is at work here, in its general form.

be no doubt you are not blame*worthy* for sneezing. To take a livelier example, suppose I am a spy overseas working with my female partner. My cover is as a married physicist; my partner playing my wife. The success of my mission depends on my ability to convince those I meet of the truth of my cover, that I am in fact the physicist I'm claiming to be. It would help my performance if I, too, believed the supposed facts of my cover. I'll be less likely to "slip up" and more likely to convince them. Thus, it would be beneficial if I were to believe that I was actually married to my partner. Indeed, I ought to believe this, as it is crucial to the success of my mission (which we can assume is of the utmost importance). Nevertheless, it would be a mistake for me to believe this claim, for it isn't true. So even if I could revise my beliefs in such a fashion, and even if I did actually revise them, I would be making a mistake. This is because the "fittingness" of belief tracks truth; beliefs "fit" when they match truths in the world. Similarly, I think blame "fits" when it is correct, when the agent is actually worthy of the blame. And this is the case so long as the agent is responsible and the outcome is bad.[166] But this result doesn't entail that we ought to actually blame him. There may be all sorts of competing reasons, just as there can be competing reasons in favor of believing some untrue claim. But what counts for the "fittingness" of blame is when the agent is liable to the response, irrespective of whether we ought to engage in it all things considered.

Rosen instead thinks that blame fits just in case we ought to engage in it. This is clear from his comments on what supports his cited norm of fairness. There what grounds the unfairness of our actual blame in the given cases is the unreasonableness of our expectations. But this is a comment on when we shouldn't blame, not on when blame

[166] Even if one rejects my particular characterization of when blameworthiness is correct, the conceptual distinction should be acknowledged.

is unwarranted. Pete Graham makes a similar point regarding the appropriateness of giving a response to an agent and the appropriateness of a particular person's giving that same response.[167] He gives the example of Ned, a notorious car thief, blaming Homer for stealing his car. We might think in such cases that it is inappropriate for Ned to blame Homer because such blame is hypocritical. Who is Ned to blame *anyone* for stealing cars? But Graham notes that just because Ned shouldn't blame Homer because it would be hypocritical, this doesn't mean that Homer isn't blameworthy for stealing the car. After all, Ned's wife Maude, who isn't a car thief, is certainly within her rights to blame Homer. Indeed, it would seem that Homer's liability to blame isn't at all affected by who does the blaming, even if we may judge that in particular cases particular individuals should refrain from engaging in blame.

These reflections on Ned and Homer's case are enlightening, for they expose part of what I think is going on in Rosen's examples.[168] Call the above the "hypocritical thought."[169] When combined with another thought, the "luck of the draw thought," we get a convincing picture of what supports the intuitions that the slave owner and sexist are not blameworthy. The "luck of the draw" thought expresses the belief that we could have been like the slave owner or sexist. Had we been born into ancient Hittite culture, we too would have likely accepted slavery without reservation. And had we grown up in the 1950's, we too would have likely come to have sexist beliefs. There is therefore something hypocritical in us blaming them since it is only due to the luck of the draw that

[167] This discussion can be found in Graham [2005], pp.173-183.

[168] Here I am in agreement with Graham, who makes similar points.

[169] Graham helpfully illustrates the thought with the trite expression "people in glass houses shouldn't throw stones".

we didn't end up like them.[170] We might be led by this reasoning, then, to suppose that it would be inappropriate or unfair for us to actually blame the slave owner and sexist.

But this reasoning is clearly inadequate for showing that the sexist or Hittite is not blameworthy. The fact that we might think it inappropriate for *us* to blame them does not show that they aren't liable to blame. For we surely do not have the same intuitions about the slave owner's fellow Hittites, or the sexist's daughters. It surely wouldn't be inappropriate for those in the respective times and cultures to do the blaming. Though it may have taken a "moral genius" to have seen the wrongness of slavery in Hittite culture, had such a genius existed, it seems entirely appropriate for him to criticize his fellow Hittite's for their immoral practices. To claim otherwise risks endorsing the view that until an immoral practice is acknowledged as such by a significant portion of the population it cannot be criticized, nor can its practitioners be appropriately blamed. So I think it eminently plausible that a Hittite slave owner could be appropriately blamed by a fellow Hittite.[171] And this observation helps support the claim that the inappropriateness of particular individuals blaming others doesn't establish that it is inappropriate *tout court* to blame others. This is precisely the result of considering Ned and Homer's case.[172]

I think the slave owner and sexist are blameworthy. Indeed, I think it is fitting for us to blame them. But I need not establish this stronger claim. All I need show is that they can be blameworthy even if it would be inappropriate for us to actually blame them. And this last fact may indeed be true. The explanation for this fact is that there can be all manner of good reasons against actually engaging in a response like blame; perhaps it

[170] As Graham notes, Watson [1987] cites similar reasoning as part of the explanation for why we hesitate to blame criminals who themselves were victims of abusive childhoods. See Graham [2005], p.181, n.96.
[171] Though not a fellow slave owner, naturally.
[172] Indeed, we can see Ned and Homer's case as illustrating a family of cases that puts further pressure on practice-based views.

would be hypocritical, or rude, or insensitive, or harmful, etc. Nevertheless, that such good reasons exist, that they might show that we shouldn't blame in a particular instance, does not show that the individual in question isn't liable to blame. In short, one can be blameworthy even if all things consider we shouldn't actually blame him.

5.2.3. Moral Knowledge

IN SPECIFYING THE No-Mistake Condition, I stated that an agent must have correct beliefs only about those non-normative facts sufficient for generating the appropriate normative evaluation. For instance, one need not believe that holding a slave is wrong, one need only believe one is holding a slave.[173] Whether holding a slave is wrong is the result of theory. Moral theory will tell us what things are permissible and impermissible, just as aesthetic theory will tell us what things are beautiful or ugly.

Nevertheless, I think that beliefs about the result of theory are unnecessary for evaluations of blameworthiness, and they certainly aren't necessary for responsibility. Indeed, knowledge of the non-moral facts certainly seems sufficient. Imagine someone responding to our censure of him with "Sure, I knew I was causing harm, but I didn't know that causing harm was wrong!" Even if he is genuinely ignorant of this fact, which seems improbable,[174] if he knows what harm is and that he's causing it, we need no other

[173] This is a bit simplistic. I think one needs to know that a slave is a human being that is treated like property, or something like this. Knowledge of a slave as the same species as the owner is part of why Rosen chooses the Hittite owner in the first place. It is unclear how the arguments here apply to slave owners in, say, the colonial American south, where many may have operated under the false factual belief that slaves were not human beings, though their subsequent treatment of slaves would still have been wrong even if their biology had been correct.

[174] There may well be individuals who do not know (or genuinely do not believe) that causing pain is wrong. My first comment is that such individuals certainly stand outside the norm and aren't the core cases we should first test moral competence against. Second, I suspect such individuals do not just believe that causing harm isn't wrong, they evince a pathology, a failed capacity to see the world in moral terms. To

information to blame him. This is surely an uncontroversial case of doing wrong; we're likely to think that any moral theory worth having will treat causing harm as usually wrong. There are no doubt more controversial cases. Suppose Sharon has an abortion (she thinks it's morally permissible). An anti-abortion group blames her for it. Is Sharon blameworthy? I don't know. All I'm arguing here is that she doesn't have an excuse simply because she thinks it morally permissible. She isn't blameworthy if it is permissible and she is blameworthy if abortion is wrong. Her blameworthiness, then, will depend on whether abortion is morally permissible or not. But we shouldn't be surprised that blameworthiness and praiseworthiness will depend on the verdicts of our true normative verdicts.

To be responsible, on my view, one need not know the verdicts of these theories, which is a plus, I think. If responsibility is dependent on moral theory, then we face two significant problems. First, it makes many responsibility ascriptions conditional. For instance, Sharon is responsible for getting her abortion, but only if abortion is permissible. Otherwise, she will have been wrong about the verdict, and thus not responsible. Now, one might argue that Sharon's case is disanalogous to the Hittite slave owner, because the claim in the Hittite's case is that it would have taken a moral genius to see the wrongness of slavery in his time. In Sharon's case, it doesn't take a genius to see how abortion might be morally wrong. But the problem is that for many intelligent moral philosophers, those who think very carefully about such problems, it isn't at all

my mind, this goes well beyond a moral competence; it suggests a cognitive deficiency of a sort. Perhaps we should then require healthy cognition for responsibility? My preferred response would be rather to hold that these sorts of individuals are responsible. But this is an argument that actually stands outside the main discussion here, for it concerns how we ought to handle special cases, and would require extensive discussion of what the right verdict in cases of such psychopathy should be. I have no settled position on this. For some helpful discussions of these issues, see, e.g., Duff [1977]; Greenspan [2003]; Watson [1987]; Wolf [1987].

clear what the right verdict is for abortion. Just because there are some people who defend both sides of the issue, doesn't mean that it won't take a moral genius to figure out the right answer.[175] But in any event, the existence of disagreement doesn't imply, to my mind, that expecting people to have the correct beliefs would be significantly more reasonable. And if we give up looking at moral competence in terms of reasonable expectations, then Sharon's case seems much more analogous to the Hittite's after all. So, if we tie responsibility to our moral theories then we're effectively hindered in rendering responsibility verdicts without the correct moral verdict (in order to evaluate whether the agent had correct moral beliefs). This is a serious problem for being able to make definitive ascriptions of responsibility; they are impossible without settled moral verdicts.

The second problem is simpler and more damaging. Tying responsibility to moral theory in this way will render counter-intuitive results. Suppose that Susan, too, has an abortion. Susan, however, believes abortion to be morally impermissible. She went through with it despite this belief (we may imagine that she had further compelling reasons). Supposing that Sharon and Susan's circumstances are sufficiently the same, and supposing that abortion in such a case is either permissible or impermissible, then either Sharon is responsible for her abortion or Susan is responsible for her abortion, *but not both*. And this seems mistaken given that they are alike in every other respect except belief in the action's permissibility.[176]

In light of these problems, I think it an advantage of any view that it rejects a normative competence condition. To be responsible one need only know those non-

[175] I happen to think abortion is this difficult an applied ethics issue.

[176] We may even suppose that Sharon has Susan's additional reasons in favor of abortion and they happen to over-determine her choice, whereas in Susan's case they outweigh the moral reason against.

normative facts sufficient for generating the relevant evaluation according to the right theory.

The Argument from Moral Ignorance fails to show that individuals must have correct beliefs about moral verdicts to be responsible. The slave owner and sexist are blameworthy, even if we have strong reasons not to blame them. Therefore, my No-Mistake Condition survives, and no amendments are necessary to my three conditions on responsibility. Nevertheless, Rosen's approach succeeds, I think, in defending grounds for a certain sort of caution about our blaming practices. While I don't think Rosen achieves the skepticism about blameworthiness that was his aim, he does show that facts about our moral knowledge invite skepticism about when we ought to go about the business of blaming. The value of our practices can be independently assessed even if we're confident, as I think we should be, that individuals are blameworthy for at least some of the things they do. Indeed, I think this skepticism is rather mild, as I don't think that there is widespread disagreement about the moral verdicts any worthwhile theory ought to produce. Nevertheless, consideration of the issues discussed above does draw our attention to thinking carefully about whether we ought to blame someone in a given situation. And increasing the amount of care and attention involved in our blaming of others is surely a positive result.

5.3. Historical Condition

MANIPULATION ARGUMENTS have been particularly popular recently.[177] They purport to show that how an agent comes to act makes a difference to his responsibility. Specifically, they indicate that the way an agent comes to have the beliefs, desires, and values on which he acts can make a difference to his responsibility for the outcome. Here I will discuss two separate illustrations of supposed responsibility-threatening manipulation, and try to show how they may be taken to illustrate an objection to my view as stated (3.1). Next, I'll take each case in turn, diagnosing what could be its problematic aspects for responsibility, but arguing in each case that we shouldn't think them worrisome after all (3.2 & 3.3). Lastly, I'll consider a different manipulation argument intended to support the view that responsibility is incompatible with determinism after all (3.4). I discuss it last because the lessons learned in the preceding sections will allow us to better appreciate my response, and help draw out a general conclusion to take away from manipulation arguments.

5.3.1. Manipulation Arguments

I BEGIN WITH TWO pairs of cases:[178]

[177] Some recent examples can be found in Fischer, et al. [2007]; Mele [2006].

[178] These cases are variations on ones given in Mele [1995, 2006]. Mele cites Kane [1985] as a source of similar sorts of manipulation examples. Mele is most interested in the conditions needed for autonomy, which he thinks is sufficient for the sort of free will necessary for responsibility, but he remains an agnostic in the debate between compatibilists and incompatibilists. He uses cases like Overnight Opera Lover in support of historical conditions on responsibility; he cites cases like Zygote Zapper as explaining why he cannot flatly endorse compatibilism. Kane uses manipulation arguments to argue directly for incompatibilism (of free will and responsibility with determinism).

Overnight Opera Lover: Donna loves the opera. She finds it musically inspired, compellingly dramatic, and a pleasure to see and hear. In addition to attending operas whenever possible, she also wants to promote her local opera company's future. As a result, she donates $1,000 dollars on June 1st to the company.

Donny, on the other hand, can't stand the opera. He thinks it repetitive and boring, he can't stand the timbre of operatic voices, and he hates the aesthetic of opera halls. All in all, he ranks operas below visits to the DMV. All his life he has petitioned for less public funding of opera companies, urging that money to go elsewhere. When Donny goes to sleep on the night of May 31st, however, he is visited by an opera-loving neuroscientist, who reconfigures Donny's brain as he sleeps. The neuroscientist has studied Donna to find out what makes her tick, and configured Donny to be her psychological "match."[179] When Donny awakes, he is the biggest opera fan in the world. Now he finds them to be musically inspired, compellingly dramatic, and a pleasure to see and hear – just like Donna. As a result, Donny donates $1,000 on June 1st to his local opera company.

Zygote Zapper: Donna loves the opera just as before. Additionally, an opera loving geneticist wants to ensure that his local opera company continues to prosper in the future, so he manipulates a particular

[179] Presumably the neuroscientist does this by manipulating neurons or other parts of the brain.

> zygote such that on June 1st, 30 years from now, the individual the zygote becomes will donate $1,000 to the company. The geneticist, too, has studied Donna to find out why she donates money to the opera. So he arranges the zygotes genetic code so as to produce Donna's psychological "match" 30 years from now. Sure enough, in 30 years, the zygote, now a man named Danny, donates $1,000 to his local opera company.

To keep things straight, here is a table illustrating the cases:

Donna	Loves the opera and donates money to support it (non-manipulated)
Donny	Has always hated the opera – but has his beliefs, desires, and values reconfigured so as to "match" Donna's psychological set – so he, too, donates money to the opera (neuronal manipulation)
Danny	Was configured as a zygote such that, in 30 years, he too is a psychological "match" of Donna – so he also donates money to the opera (genetic manipulation)

In the cases above, we are meant to think that neither Donny nor Danny is responsible for donating the money. Each has been manipulated and "made" to donate the money. But in both cases the agent in question seems to meet my conditions on responsibility. They each bring about the donation voluntarily, intentionally, and without mistake. Each one's action is certainly explainable by a belief-desire set. Moreover, we can easily cite the beliefs and desires that explain the action. Each wants the opera company success in the future and believes that donating money will help ensure some measure of financial stability. And they are both correct about the relevant features of their donations; they are, in fact, giving money to a local opera company that will use it to help fund future projects. So, if Donny and Danny aren't responsible for their donations, then I'm missing a condition on responsibility, and so my three aren't sufficient by themselves.

The above cases are typically used to motivate an argument for a *historical condition* on responsibility. Not only must one act in a particular way in order to be responsible, one must come to act in a particular way. Specifically, one must come to act in a manner free of problematic manipulation, as seen in Donny and Danny's cases. The problem in these cases is that the beliefs and desires and values doing the explanatory work are, in some significant sense, not really Donny's and Danny's. At least, they haven't come to have those beliefs, or endorse those values, in the appropriate way. This deviant history lies behind the claim that neither Donny nor Danny is responsible. Whether or not we can clearly demarcate between deviant and non-deviant histories, it seems clear, according to manipulation arguments, that Donny and Danny are not examples of normal belief acquisition or normal value endorsement. Something has gone wrong in the histories leading up to the donations, and this deviant history explains their undermined responsibility.

According to my view, Donny and Danny are both responsible; they satisfy my three conditions on responsibility at the time of their donation. And I think this is the right view to take about these cases. I'll argue as much, first tackling Donny's case, and then Danny's. My view, then, is a non-historical one; it doesn't require that one's history leading up to an action satisfy some condition in order to be responsible for the outcomes. I'll say more about the non-historical dimension and its relation to responsibility for how one is a bit later (3.4).

5.3.2. Overnight Opera Lover

LET'S BEGIN WITH SOME simple observations about Donny's situation. Let's take the perspective of one of his close friends. In fact, the friend and Donny started P.A.P.F.O (People Against the Public Funding of Opera) together. They've always shared a disdain for opera; and an interest in a great many other things, like fishing, video games, supermodels, and beer. In fact, they had just gone fishing on the 31st of May, and had arranged to go to the P.A.P.F.O meeting together the next day. When Donny's friend arrives, however, Donny says that he's not going to the meeting. This strikes Donny's friend as strange, and he asks why. Donny says he now loves the opera. It's a fantastic art form. In fact, he's quitting P.A.P.F.O altogether. Now Donny's friend likely thinks Donny is just joking with him. Maybe he laughs nervously. But when he sees Donny is serious, he's liable to say something like, "What's come over you? You're not yourself today!" Donny is unaware of how the change has taken place, though he is aware of a change. So he's likely to reply that he woke up this morning with a new (though inexplicable, to him) appreciation for opera and a strong desire to promote its continuation. In fact, he says proudly, he's headed right now to donate $1,000 dollars to the local opera company.

No doubt, Donny' friend is flabbergasted. He doesn't know what to make of this abrupt and sudden change in Donny. He's known Donny for years, and his behavior this morning isn't like him at all. Indeed, when Donny's friend goes to the P.A.P.F.O meeting and informs them of Donny's change, the entire group is likely to be dumbfounded. How strange for someone to undergo a complete reversal of such a strongly held conviction. To his closest associates, Donny will likely seem a different person.

At least one significant feature of the case, then, is the extreme suddenness of the change. Compare Donny's case as told to an elaborated version of Donna's story. Suppose she didn't always love the opera. For many years she thought it stuffy and boring. But then she dated someone who loved the opera, and so she actually went to one. And, much to her surprise, she found it wasn't nearly as bad. She attributed her previous dislike to growing up in a smallish city and not being exposed to 'good' opera. She is interested in learning more and so begins to develop an appreciation for opera. Her appreciation grows; so much so that in a matter of months she finds herself donating $1,000 dollars to the local opera company. Donna's friends may also be surprised at the change (it is still a quick development), but it is gradual enough not to strike them as bizarre.

Moreover, unlike Donny, Donna can explain where this new love of opera originated. She can recount for her friends the sources of her altered beliefs and, roughly, the process by which she came to change her outlook. Donny's change is inexplicable, even to himself (we, of course, know differently). And this brings us to the second significant feature. The neuroscientist is clearly "behind" Donny's change. It is his plan to "make" someone like the opera, and he carries out his plan using Donny. This aspect of Donny's case no doubt lies behind much of what concerns us about his actions. He isn't really responsible for the donation, because the neuroscientist is responsible for the way Donny is. I will call agents like the neuroscientist in this case *Intervening Immune Agents*. They're intervening because they influence in significant ways what the other agents in the scenario do, and they are immune because they are not subject to the same sort of influence. In Donny's case, the neuroscientist is the one who reconfigures his psychological set to look like Donna's (at least with respect to opera), but the neuroscientist himself isn't subject to such psychological tampering (by hypothesis).

The third significant feature is importantly related to the second. I've already mentioned it in fact. Part of why Donna can recount her change is that it came about through the usual process. As agents, we have the capacity to reflect on our beliefs, desires, and values. And while in most circumstances we cannot spontaneously revise them, we can take steps so as to revise them or to encourage the adoption of different values. Donny's change doesn't come about through this usual route. His change is engineered by another party, who bypasses his reflective capacities and simply puts a particular set of psychological states in his head. So while we may be comfortable with the claim that Donna exerts some control over her psychological set, at least in coming to be a lover of opera, we cannot make such a claim of Donny. He exercised the same sort of control in coming to be a hater of opera, but the neuroscientist's intervention lies behind his newfound appreciation, and this process bypassed the usual routes.

I think these three features are significant in that they form the bulk of support for the judgment that Donna is responsible for her donation while Donny is not responsible for his. First, the change is extremely sudden (the Suddenness feature). Second, the change is primarily the result of an Intervening Immune Agent (the IIA feature). And third, the change bypasses the agent's reflective capacities (the Bypass feature). I now take up these features in turn, arguing that they should not lead us to think that Donny is not responsible after all.

5.3.2.1. Suddenness

DONNY UNDERGOES HIS change in values extremely suddenly. And this is a very important fact; for it immediately directs our attention to the way in which Donny's case is aberrant. In general, people don't undergo such rapid and radical changes in their psychological sets. In general, people tend to revise their core convictions through gradual processes. They don't embrace radically different conclusions overnight.

But sometimes they do. Sometimes individuals experience an "awakening" or sudden realization that immediately shapes their lives in particular ways. I'm thinking specifically of individuals who experience epiphanies.[180] The change can be quite dramatic and quite sudden. Indeed, when faced with such individuals we often confess that they seem to be an entirely different person. Nevertheless, I take it we don't think in such circumstances that the individual isn't responsible for the things he does, even those things he does shortly after his change. The born-again Christian who immediately begins donating to charities and being more considerate of others is surely praiseworthy for such deeds. And, I would think, the same could be said about a born-again "Satanist," who immediately begins doing ill work about town; he seems blameworthy for such acts.

Moreover, I should think that we want to be responsible for such acts. We should allow for the possibility of quick changes to our psychological sets. Imagine that Sasha has been operating under some false belief, say, that eating meat is morally permissible. Sasha has even reaffirmed her commitment to its permissibility over and over again, even in the face of good philosophical arguments to the contrary. But then suppose that Sasha comes across a new argument for vegetarianism. It is so new and so good that it

[180] For example, some religious converts who are "born again" or "see the light" quite suddenly.

convinces her that eating meat is really morally *im*permissible. She immediately stops eating meat, and is glad that she can take credit for her new acts of compassion, even as she might regret her past meat-eating acts, committed as they were under an error. Nevertheless, I should think that once we've come to the new belief, we should want immediate credit for doing what's right.[181]

So, I don't think it can be the suddenness of the change that explains why Donny isn't responsible for the donation. There are plenty of examples, it would seem, where the change is just as radical, just as sudden, and yet we think responsibility remains intact. Perhaps the suddenness feature attunes our attention; it raises our suspicions, as it were. We begin to look more closely for explanations of the change. But I don't think that by itself it can show responsibility to be undermined.

5.3.2.2. Intervening Immune Agents

THOUGHT EXPERIMENTS WITH intervening immune agents are all too common in the responsibility literature. In Donny's case, there is a neuroscientist who reconfigures Donny's psychological set. When Donny makes his donation, the existence of the neuroscientist is extremely important. It is easy to say that the neuroscientist is responsible for the donation. After all, he altered Donny's brain specifically with the purpose of getting him to support his opera company. Donny seems to be an instrument of the neuroscientist's purposes; a puppet whose strings the neuroscientist pulls.

[181] I am supposing here that eating meat is morally problematic, but we could alter the case to use a different moral claim.

But I think we have the same reaction, though perhaps to a lesser degree, to similar cases that nonetheless do not seem to involve undermined responsibility. Take a case of manipulation as found in Shakespeare's *Othello*. Iago manipulates Othello into believing that Desdemona has been unfaithful, and, in a grossly simplified summary, Othello kills Desdemona. Many who have read the play no doubt revolt at Iago's role in the affair. He is the mastermind behind the plot, and many of the events transpire as they do due to his actions, shaping others' beliefs through deception and stoking their desires. Still, it strains me to believe that Othello isn't responsible for Desdemona's death. Though he is in error about his reasons for killing her, he kills Desdemona intentionally, in full awareness. It is his actions that lead most directly to her death, and while he's been fed false information, he brings about her death by way of action that would in all other circumstances be the paradigm for responsible action.

Now, it seems to me there are two likely objections to my drawing this parallel. The first objection is that even if we grant Othello is responsible for Desdemona's death, he is only *partially* responsible. After all, he and Iago share the responsibility for her death. The picture such a claim as this makes is that there is some amount of responsibility available for the outcome in question, to be apportioned out to those who took part in the plot.

But this doesn't seem right. After all, we might widen our gaze to include all those characters (such as Roderigo) that joined in Iago's scheme and those who were unwitting accessories (perhaps like Emilia).[182] But the more people we add to the story does not reduce the responsibility apportioned to any one of them. Just as a bank robbery

[182] For those who aren't familiar with the play, Roderigo enlists Iago's help to win Desdemona's heart, and Emilia steals Desdemona's handkerchief for Iago, one of the pieces of material "evidence" he uses to deceive Othello.

could be committed by a team of seven or seventy people, but we don't suppose that one of the seven is more responsible than one of the seventy, we shouldn't think that Othello's responsibility diminishes just because Iago is implicated.[183] For one thing, neither the bank robbers nor Othello seem any less blameworthy, no matter the number of people added to the story. Moreover, while Iago certainly contributes to the beliefs and desires Othello acts on in killing her, it would seem that her death is primarily Othello's doing.

The second objection picks up on this line of thinking. It suggests that while her death may be primarily Othello's doing, Iago nonetheless plays a crucial role. Without his interference, Othello would likely not have come to believe Desdemona was having an affair. So, without Iago's meddling, Othello would likely not have killed her. So, Othello's killing of Desdemona is *less* his doing than had he come to have the same beliefs of his own accord (whether these were true or not).

But again, this doesn't seem right. For suppose that Sasha came to her new vegetarian beliefs by way of argument by Sean. He wanted her to "see the light" regarding respecting animals, and he persuaded her through meticulous and creative argumentation. In fact, he wanted her to stop eating meat, and this why he said the things he said.[184] Her selection of a vegetarian offering and her next meal is her doing, but we might plausibly insist that it is partially Sean's doing as well. We can, if necessary, further suppose that he went out of his way to convince her, providing her with documentation about food animal raising practices and the like. Despite all his work,

[183] Frankfurt makes a similar claim about aggregation in Frankfurt [1971]. My thanks again to Manuel Vargas for bringing this to my attention.

[184] I actually think it is immaterial whether or not Sean's argument is sound, and not just simply valid. Even so, we can proceed here under the assumption that his argument was both valid and sound.

however, and even if he is partly responsible for her vegetarian selection, I don't think this in any way diminishes Sasha's responsibility for the selection.

In fact, I think the better answer is that Sean is partly responsible for how Sasha *is*. He is partly responsible for her beliefs about meat eating. After all, he shone a new light on them through documentation and rational argument. Similarly, Iago is partly responsible for how Othello is. And, I think, this is clearest in Donny case. The neuroscientist is largely responsible for how Donny is; probably entirely responsible for how he is regarding opera. In all these cases, however, I still think that the agent in question is responsible for the relevant outcome, even if we can also implicate another. What seems to distinguish Donny's case is not that there is another agent (the neuroscientist) implicated in his psychological change, but the way in which the neuroscientist brings about the change. In other words, it is the deviance from our standard ways of coming to believe new things that makes the difference in Donny's case.

5.3.2.3. Bypassing

SASHA'S BELIEFS ARE CHANGED through rational argument and new evidence. We change our beliefs often enough by such a process. It is well within the norm for psychological revision. But neuronal alteration stands far outside the norm. This is what makes Donny's case different. He is psychologically altered, in ways similar to Sasha and Othello, but his manipulation seems much worse, I submit, because it bypasses the usual route to such changes. Donny is unaware of how this change comes place, whereas both Sasha and Othello can explain at least what new evidence they were given and why it changed their minds. The picture we get in their cases is one in which the manipulating (or persuading)[185] agent submits evidence for the other's consideration. So it's up to Sasha and Othello whether to entertain the new evidence, and then it's up to them whether or not to revise their beliefs accordingly. The neuroscientist, on the other hand, directly alters Donny's brain, and he doesn't get to consider anything. In Donny's case we're more likely to say he's been used or violated, in large part because of this bypassing.

Still, I'm not sure that bypassing is sufficient for undermining his responsibility. Let's assume that the neuroscientist is responsible for the way Donny is. This I think is a common explanation to (at least virtually) all cases of manipulation of the sort we're interested in here. Granting that assumption, is the fact that this change resulted from bypassing the normal conscious routes to belief revision sufficient for undermining Donny's responsibility for his donation? I don't think so. I think bypassing of this sort is only worrisome if we retain an implausible notion of our conscious belief revision.

Sasha is supposed to be a case of non-bypassed belief revision. She comes to rethink her position on eating meat, and changes her mind, believing now that it is immoral. Similarly, Donna used to hate the opera, believing it to be stuffy and boring, but now she likes it very much. She now believes that it is artistically inspiring and compellingly dramatic. She too is an example of non-bypassed belief revision. Now, in order for it to be the bypassing that undermines Donny's responsibility, it isn't enough to simply be able to draw a line between his case and theirs. One also needs to show that the line distinguishes his case in an important and interesting respect from the standpoint

[185] I use persuasion to talk of Sasha cases, where we don't typically think Sean does anything untoward in convincing her. I think the differences between manipulation and persuasion are interesting and merit additional discussion, but I won't pursue them further here.

of belief formation and revision. In particular, it seems that the standard cases appeal to the notion of the agent having control over his attitude revision, and that it is precisely this control that Donny lacks.

I grant that we can distinguish between Donny's case, on the one hand, and Sasha and Donna's case on the other. Donny cannot explain how his change in attitudes can about, whereas the women can. And this feature seems to show Donny's case to be non-standard, since we usually can explain what led us to change our minds. But I don't think this feature makes Donny's case different *enough*. I suggested earlier that the picture we get of standard cases like Sasha's and Donna's is one where the agent is given evidence to consider and then it's up to them whether to accept it and revise their beliefs accordingly, or reject the evidence altogether. Such a picture makes it seem that it's up to us whether or not to change our attitudes. But I'm skeptical that much of our mental life is so directly under our control. My position is consistent with the claim that we can sometimes believe at will, so long as most of our beliefs and other mental attitudes are not under our control in this way.[186]

When I have evidence that, say, there's a computer screen in front of me when I can see it there, I cannot help but believe that it *is* there (barring other beliefs that might weaken such visual evidence). Similarly, if I am given a valid bit of reasoning, whose premises I believe to be true, I cannot help but accept the conclusion. I may *resist* the conclusion; I may look for flaws in the argument, or try to find reasons against accepting

[186] There is a vast literature on believing at will, most of the discussions center on its very possibility. While I'm skeptical of such a claim, it is not my focus here, nor do I require believing at will to be impossible in order to mount my objection. Rather, it just has to be the case that we rarely do believe at will; specifically, that Donna and Sasha do not believe at will, but rather in a significant sense, do so involuntarily (though the sense in which this is so will obviously differ from my use of involuntary actions throughout this dissertation). For some of the relevant arguments on the possibility of believing at will, see Williams [1973]; Naylor [1985]; and Winters [1979]. For a discussion of the possibility of desiring at will, see Shemmer [2004].

it, or insist that it isn't true. But all of these are efforts on my part to find a way so as not to believe it. By hypothesis, Sasha takes the argument for vegetarianism to be sound. She, I submit, cannot but accept the conclusion. She is forced by rational pressure to do so. I suppose we might allow that she can fail to accept the conclusion, but only on pain of gross irrationality. We would think something has gone seriously wrong with her reason, or that she is being disingenuous. And in most cases of apparent irrationality, I suspect that Sasha would still believe the conclusion reluctantly, though she might fail to assent to its truth. But no one would deny that we have significant control over our speech acts.

I don't think Sasha can but believe in the conclusion that eating meat is immoral. Nor do I think that she can choose whether or not to accept each premise of the argument. She can perhaps choose to weight the evidence in favor of each premise to some degree. As Naylor notes, we can "rethink our assumptions about what counts as evidence."[187] And so perhaps it is possible to deceive ourselves into believing some proposition by managing the evidence for that proposition in various ways. But this isn't how things usually go. In standard cases, we find ourselves presented with evidence that either supports the proposition or doesn't, and we find ourselves either compelled to accept it or unconvinced.

Donna comes to her attitudes about opera even more gradually than does Sasha. But even here, I don't think she exerts that much control. Remember she first reencounters the opera because she's taken there on a date. She certainly chose to go out on the date; but once we examine the reasons many of us choose who we date, I think we find a laundry list of factors typically outside our control. We can embellish Donna's

[187] Naylor [1985], p.434.

story to observe the point. She is typically attracted to men who dress a certain way, who drive a certain type of car. She isn’t sure why this is the case, and she has often thought that it might be better to widen her scope. She’s even endeavored in this direction, agreeing to dates from guys she isn’t attracted to in order to try and combat her bias. She has at times lamented her “type” and wished it were different. Nevertheless, she *is* often attracted to guys of a certain type, and it is just such a guy who takes her to the opera. As he drives her to the opera house, Donna is dreading the evening. (Remember, at this point she still hates the opera, and she is sure she’s going to have a terrible time). Though her date is regaling her with the many finer points of operas (as he considers himself quite the aficionado), she is highly skeptical (at least to herself). As they are led to their seats, she is imagining all the ways she can cut the night short, to try and find ways to get out of having to withstand the entire opera. But as the performance begins, she is amazed. Much to her chagrin, she finds herself enthralled by the costumes and the sets, the voices are rich and impassioned, and the score is truly outstanding. Throughout the performance, she is honestly surprised to be having such a good time. In the car ride after the performance, she admits as much to her date, and speaks freely about how she had previously disliked the opera.

Now Donna does take active steps to encourage her newfound appreciation for opera. She goes to the library and gets books to further her understanding, she buys recordings of famous operas, and she attends more live performances. But the ways in which the performances and recordings strike her are not really under her control. After all, she was planning to have a bad time on her opera date. And we can imagine that the second time she goes she again expects to have a bad time, thinking the last experience to be a fluke. But she is surprised again to thoroughly enjoy herself. And her experiences at these performances contribute significantly to her changing her attitudes toward the opera. They are, I think, the most significant contributing factor.

I think Donna's case helps illustrate that while one can attempt to direct one's attitudes, or to take steps to modify or limit their influence, especially when one considers them to be problematic (as in the case of certain prejudices), we do not have the sort of control over our attitudes we might think we have.[188] And if we don't have such control, we might not think that in cases in which the agent's reflective consciousness is bypassed responsibility is undermined. We can still distinguish between Sasha and Donna, on the one hand, and Donny, on the other. Donny's brain is manipulated by the neuroscientist, and he can't explain how his new attitudes came to be. Nevertheless, it doesn't seem as though he has radically less control over his attitudes than Sasha and Donna, because it isn't clear that they have much control over theirs.

Obviously, this doesn't mean the differences in Donny's case aren't important. Quite the contrary. The neuroscientist violates Donny in a very serious manner, changing core aspects of his personality. Donny seems a different person, and this result was brought about by the neuroscientist's plan. The neuroscientist is responsible for the way Donny is, for "messing around" with his psychological set, for invading his brain to make him into a lover of opera. Sasha's interlocutor and Donna's date are not invasive at all. What the neuroscientist did is unimaginable, horribly wrong, and worthy of our harshest reprisals. Donny has been changed, without his consent or knowledge, into someone who holds attitudes radically opposed to those he had before. We can condemn such

[188] Again, this doesn't support the stronger claim that it is impossible to believe at will, or to directly influence our attitudes. I only claim that we don't typically revise our attitudes with such control.

alterations, even in cases in which the newly acquired attitudes are superior to those previously held; for example, if the neuroscientist had changed a racial bigot into a more tolerant individual. As I mentioned above, the role of the Intervening Immune Agent plays a significant role in how we see Donny's case. But even though he bypasses the usual routes to attitudes adjustment, I don't think that these routes matter in distinguishing Donny from Sasha and Donna in terms of their responsibility for their donations.

I think that even bypassing fails to show why Donny isn't responsible for his donation. Donna and Sasha come by their new attitudes in the standard manner, but this way seems to lack significant control on their parts, just as in Donny's case. While there are still important differences between the cases, I think the differences trade on features irrelevant to Donny's *responsibility* for his donation to the opera company. Donny's case is unfortunate; it implies that it is possible that some devious agent could change us so as to make us responsible for what he wanted us to do. I think this is right, partly because in persuasion and non-undermining manipulation cases the same general story is true, and partly because I think this feature supports why we think so badly of the neuroscientist.

While the neuroscientist is likely responsible for how Donny is, Donny is responsible for what he does. Those who know him will likely think him a significantly different person because his change is so radical. His sudden reversal will strike them as strange, but this is true in cases of attitude revisions that don't undermine responsibility. The neuroscientist's involvement as an Intervening Immune Agent skews our intuitions, since he plausibly shares responsibility for Donny's actions by altering Donny with the intention of bringing those outcomes about. He is also primarily responsible for how Donny is psychologically. But so are the intervening agents in Sasha and Othello's cases. Indeed, not even the fact that the neuroscientist's meddling deviates from standard cases by bypassing Donny's psychology seems to establish the necessary difference.

I conclude that Overnight Opera Lover as a case of manipulation fails to establish that Donny's responsibility for his donation is undermined, and thus fails to threaten my conditions for responsibility as jointly sufficient. But it is an important example nonetheless. It shows that we have less control than we might have thought over our psychological sets, and that those who influence those sets can play an important role in what we ultimately do. Still, I do not think that Overnight Opera Lover forces us to abandon my conditions on responsibility. Perhaps Zygote Zapper can do better.

5.3.3. Zygote Zapper

RECALL THE DETAILS of the case:

Zygote Zapper: Donna loves the opera just as before. Additionally, an opera loving geneticist wants to ensure that his local opera company continues to prosper in the future, so he manipulates a particular zygote such that on June 1st, 30 years from now, the individual the zygote becomes will donate $1,000 to the company. The geneticist, too, has studied Donna to find out why she donates money to the opera. So he arranges the zygotes genetic code so as to produce Donna's psychological "match" 30 years from now.

> Sure enough, in 30 years, the zygote, now a man named Danny, donates $1,000 to his local opera company.[189]

We are supposed to find that Danny isn’t responsible for his donation, as a result of the geneticist’s manipulation. In many ways, Danny’s case is harder to diagnose than Donny’s. It’s more difficult, I think, to specify how he comes to love the opera, since I, at least, am fuzzy on just how genetic manipulation can lead to a specific action 30 years later. Obviously, operating in the background is some notion of determinism, that given the facts as some point in time and the laws of nature, every truth after that point in time is guaranteed. This is necessary for the argument, since it must be the case that Danny necessarily comes to have the same psychological set as Donna, simply as a result of his genes. We needn’t assume, however, that Danny was determined to donate the money, though this point will not affect the argument, I think, nor my rebuttal.

My response is in many ways more simple than in the previous section. First, I do not think that the suddenness feature or bypassing feature is present in this case. Danny comes to love the opera at least as gradually as Donna does, in fact we might imagine he does so in a very similar way. So similar, in fact, that his conscious mind isn’t bypassed in the course of his attitudes being revised. So it must be other features of this case that are doing the work. I suspect it is the IIA feature, which leads me to the second reason my response is shorter here. I will not repeat my concerns about the Intervening Immune Agent. I will simply note at the outset that the geneticist seeks to secure a donation for the opera company through his manipulation of Danny. This is all I need to draw out

[189] If the reader finds it implausible to suppose that a geneticist could be able to manipulate genes so as to guarantee a particular result, the reader is free to alter the example, for instance, so that instead of a geneticist, it is an all-knowing, all-powerful being that does the genetic manipulation.

why the geneticist's inclusion is problematic if the thought experiment is to succeed. I do not think that the geneticist's intervention should lead us to think Danny's responsibility is undermined unless we already think that incompatibilism is true. That is, when properly diagnosed, Zygote Zapper should persuade only those who are antecedently committed to incompatibilism. But incompatibilists do not need to Zygote Zapper; they believe Danny is not responsible anyhow. So Zygote Zapper does little to convince those who are not already convinced of its conclusion. To put it another way, Zygote Zapper is only as convincing as our incompatibilist intuitions are. To the extent that one rejects incompatibilist intuitions, Zygote Zapper fails.

I think the most salient feature of Danny's case, the one the tugs at our intuitions the most, is that Danny's life seems preplanned by the geneticist. A major (we may suppose) facet of his life, his love of opera, was "implanted" in him in the genetic level. He seems to be a product of genetic tampering, not the "source" of his passions or the decider of what he values. I've already suggested in my discussion of Overnight Opera Lover that our evaluations of the amount of control we have over deciding such things is likely over-inflated; nonetheless, there is something true about the claim that Danny's life is unduly influenced by the geneticist. The geneticist determines a significant fact about Danny, his love of opera, and does so for specific ends. The geneticist loves the opera and wants to secure its future support, and so he uses Danny, without Danny's knowledge or consent, to help achieve that aim.[190]

[190] The geneticist could have very different aims, of course. He could have engineered Danny "on a whim," or because he likes the idea of making creatures who do very specific things. These points do not affect the example, the argument, or my reply.

While the manner of the manipulation is most definitely different, Danny's case seems sufficiently similar to Othello's. Indeed, the geneticist manipulates the beliefs, desires, and values that Danny has to the same effect as Iago manipulates Othello's. The way in which he does it is different, but the result is the same. And so too, I think, are many of our reactions to the case. The geneticist is responsible for how Danny is psychologically, and perhaps partially responsible for the donation. He is also guilty of manipulation, though not of the deceptive sort Iago is. But similar to the neuroscientist in Donny's case, the geneticist tampers with features central to Danny (his genes), and violates him in a most egregious manner. And as was the case with Donny, while I think this is more than enough to put the geneticist on the hook for Danny's being the way he is (at least with respect to his love of opera) and partially for what Danny does, it does not get Danny off the hook for what Danny does. This line of response marshals the same resources as in the previous section. I will rehearse this line of response no further.

Instead, I want to consider a slightly different response, though it admittedly builds off of my comments about the Intervening Immune Agent. Let's compare Danny to Donna. Both come to love the opera. Both were created with a particular genetic makeup, and both have been subject to the same laws of nature. Let's also assume that determinism is true.[191] If that's the case, then there is even less difference between Danny and Donna. Both have genetic makeups that guarantee they'll come to love the opera, and donate money to their local opera companies. If this is the case, then only the

[191] Only libertarians, incompatibilists that believe determinism is false, will balk at this assumption. But recall (from the Introduction chapter) that I don't consider libertarianism much in this dissertation because its conditions on responsibility, whatever they are, will be stronger than those I lay out here. As such, I think their arguments will have to be tougher to demonstrate and likely subject to more objections than mine here. In other words, libertarianism works harder than I think it needs to in order to establish responsibility.

role of the geneticist can make a difference. Now there are two things to say. First, if determinism is true, than his genetic manipulation of Danny is also determined by the way he is (as I'm treating it here, the past) and the laws of nature. This is a complicating wrinkle that I think can be ignored. Second, if I'm right that the mere inclusion of the geneticist does not suffice to undermine Danny's responsibility, then we can effectively eliminate him from the scenario. But if we do that, then Danny's situation looks identical to Donna's. Both are born with a certain genetic makeup, grow to love the opera, and donate $1,000 dollars as a result. But if Donna's case doesn't arouse our incompatibilist intuitions, I don't think that Danny's case should. The geneticist's role is inconsequential to Danny's responsibility, and thus his responsibility is only undermined if his action's being determined undermines responsibility. One is only committed to this claim, however, if one thinks incompatibilism is true.[192] We can emphasize this point by tweaking Danny's case just a bit. Suppose that instead of a geneticist (or supreme being) manipulating his genes, suppose a random photon passes through the zygote that becomes Danny, altering his genes in a way that determines he will donate the money. Here, I think, we should be much less inclined to judge his responsibility undermined, for, if determinism is true, there will always be some "just so" story for why someone did something. More to our purposes here, Donna may have come to love the opera in part because of circumstances *in utero* that determined she would be that way. But if compatibilists are not worried about causal determinism independent of manipulation arguments, I see no reason they should be within such arguments.

[192] Mele makes the same observation. See Mele [2006], pp. 189-192. But he suggests a better measure of intuitions would be to see what "reflective agnostics," people who have thought long and hard about responsibility but remain agnostic about the correct position to take, think about zygote manipulation cases. Nevertheless, Mele admits that arguments like Zygote Zapper are insufficient for convincing him that compatibilism fails.

Compatibilists, therefore, need not worry about manipulation arguments of Zygote Zapper's type. These arguments themselves require the truth of incompatibililsm in order to establish undermined responsibility. Moreover, as a general rule, we shouldn't rely on arguments about when responsibility is undermined that depend upon assuming the conditions on when responsibility is undermined. In other words, Zygote Zapper and its ilk only succeed if we already think incompatibilism is true; they don't give us additional reasons for its adoption. By itself, Zygote Zapper can't establish undermined responsibility in Danny's case, thus it too fails to show that the set of my conditions is insufficient.

5.3.4. Pereboom's Four-Case Argument

DERK PEREBOOM HAS given an abductive argument for incompatibilism that involves responsibility-undermining manipulation.[193] The general structure of the argument is as follows: Pereboom purports to give us a case of responsibility-undermining manipulation, but in which the agent otherwise satisfies any compatibilist conditions on responsibility you like. Next, he gives us a diagnosis of why the manipulation undermines responsibility. It is because the agent's action was causally determined by factors beyond his control. But Pereboom goes further. He then present two further cases in which he thinks there is responsibility-undermining manipulation, but of a weaker sort than in the first case, and in which, again, the agent satisfies all extant compatibilist conditions on responsibility. He suggests that it is causal determinism doing the undermining work in these cases too. Pereboom concludes his argument with a normal case of action in a

[193] The latest presentation of this argument, and some illuminating discussion of it, can be found in Fischer, et al. [2007], esp. pp. 93-101.

deterministic world, and challenges the compatibilist to distinguish between this case and at least one of the previous three. If the compatibilist cannot, Pereboom claims, than it would seem that responsibility really is incompatible with determinism. If determinism really is the best explanation for the undermined responsibility in the cases of manipulation, then we can generalize from these cases to an incompatibilist conclusion: determinism must undermine responsibility everywhere.

I don't think causal determinism is the best explanation for the intuition that the agent in the manipulation cases is not responsible. Nor do I think that Pereboom's generalization strategy works here. I'll elaborate on both these point shortly. First, I want to briefly present the four cases.[194] In each case, Prof. Plum decides to attack Ms. White with the intention of severely injuring her for some personal gain. He does so satisfying all compatibilist conditions on responsibility. His action is caused by desires that flow from a stable character (Hume), the desire conforms with his higher-order desires (Frankfurt), and he is receptive and reactive to the relevant pattern of moral reasons (Fischer & Ravizza).[195]

Case 1: Professor Plum was created by neuroscientists, who can directly manipulate him by radio-like technology. They manipulate him to

[194] I've modified the cases only slightly from Pereboom's presentation and the details I've changed do not affect the argument at all. Instead of Plum killing White, as in Pereboom's examples, in my cases Plum severely beats White, but White survives. I find it unfortunate that so many examples in the responsibility literature involve agents killing each other. No doubt this is to invoke the strong moral judgments we have when faced with murder. Nevertheless, I don't think our judgments are much weaker when it comes to aggravated assaults. And so I've taken pains throughout this dissertation to limit my examples to cases of severe harm, but short of death, wherever possible. I do the same here.

[195] Pereboom takes pain to insure that Plum meets all the relevant compatibilist criteria in his cases, and even if he doesn't, I'm confident similar cases could be constructed that did. I'll discuss this aspect of the argument no further.

reason in a particular way so as to bring about his particular desires, which in turn cause him to attack and injure White.

Case 2: Plum is like ordinary humans, except neuroscientists programmed him at birth to weigh and act on reasons in an extremely egoistic way. As a result, in his current situation, he is causally determined to deliberate in the particular way he does, with the determined consequence being his attacking and injuring White.

Case 3: Plum is like ordinary humans, except he was conditioned from birth by rigorous training practices of his community to be incredibly egoistic in his reasoning. He was too young to prevent or alter this training from determining his egoistic character, which now determines him to reason egoistically, and causes him to attack and injure White.

Case 4: Determinism is true, and every event is completely causally determined. Plum is an ordinary human, raised in normal circumstances, who is extremely egoistic. He reasons normally and attacks and injures White.

With these cases before us, Pereboom claims that it is Plum's action being causally determined by factors outside his control in Case 1 that undermines his responsibility.

Likewise, it is his being causally determined in Cases 2 & 3 that undermine responsibility there, and so in Case 4, since there is no relevant difference between it and Case 3, responsibility must be undermined as well. Thus, responsibility is in fact incompatible with determinism.

I disagree. First, I want to challenge Pereboom's abductive claim that it is determinism doing the undermining work. As Mele has noted,[196] if one thinks that Plum's responsibility is undermined in Cases 1-3, this thought is unlikely to change if the manipulation has only a 99% chance of success. So it seems unlikely that it is determinism doing the work, since intuitions are likely to be preserved even in indeterministic scenarios. Moreover, as I have argued above, there are a number of other plausible considerations that may be doing work in generating judgments of undermined responsibility in manipulation cases. And these considerations seem to be present in Pereboom's cases as well. Pereboom's abductive move seems to me to fail as a result, and the generalization to Case 4 fails as well.

Second, however, I think there's a larger problem with Pereboom's strategy. Pereboom admits that he's giving a generalization argument. So, he claims that here's a case of undermined responsibility (the neurological manipulation in Case 1), and that it is such a case because of causal determinism (his abductive move). Well, now, if that is the right diagnosis, then of course incompatibilism follows. But that's *not* how generalization arguments are supposed to work.[197] They are supposed to work by pointing to an independent condition that would globally obtain if causal determinism

[196] Mele [2006], pp. 138-144.

[197] For a wonderful discussion of generalization arguments, see Wallace [1994].

were true.[198] If one's purported condition was "causal determinism," the structure of the generalization argument falls flat, especially because compatibilists are under no pull to accept it as the explanatory condition in the first place. The force of generalization arguments is that we can generalize from an *uncontroversial* undermining feature (not a case) to a claim that determinism would globally instantiate such a feature. If determinism were true, such arguments conclude, then the all of our actions would have this uncontroversial (or at least strongly supported) undermining feature as well. Pereboom's argument doesn't do this, instead it argues that if causal determinism undermines responsibility in a given case, it does so in all other cases. But this isn't a terribly impressive argument, especially against compatibilists, since they can simply reject the antecedent. This is much easier to do in Pereboom's case, since the antecedent is precisely what the disagreement between the two camps is about. And Pereboom has given us no other reason other than intuition-pumping supporting the claim that it's causal determinism doing the work (and Mele's objection has severely weakened the plausibility of this claim). Compatibilists are likely to make different judgments anyhow, so Pereboom's Four-Case Argument does nothing to advance the dialectic. Compatibilists are always free (pardon the pun) to argue the other way, claiming that Plum is responsible in Case 4, so he is responsible all the way back to Case 1.[199] This is my preferred position since I don't think manipulation undermines responsibility. But, more importantly, Compatibilists are not forced to do this, since it is up to them to give a

[198] This is the move made by incompatibilists like van Inwagen, who argue that if determinism were true, we wouldn't have the ability to do otherwise. Most indirect arguments for incompatibilism are generalization arguments.

[199] Indeed, we might think there is a line that divides Case 1 from the others. For instance, we might think that the continuous and invasive control of Plum by the neuroscientists at each and every step render Plum a non-agent. And we might think that he is at least an agent in the other three cases. In a personal exchange, Mele has suggested the same point to me.

different explanation for why manipulation of the sort involved in Cases 1-3 undermines responsibility.

For these reasons, I prefer Mele's manipulation arguments, since they are directed at compatibilists from an agnostic position, and argue that there is something about manipulation that requires compatibilists to revise their views (at least insofar as demanding the addition of a historical condition on responsibility). I've shown why I think these arguments fail. Thus, manipulation does not require any amendments of my conditions on responsibility.

5.4. Ultimate Responsibility

WE MIGHT THINK THAT responsibility for what we do depends on responsibility for how we are (mentally). We might think that the reason Donny isn't responsible for what he does is because he's not responsible for the way he is (his beliefs, desires, values, etc.); the neuroscientist is responsible for that. And we might further think that since what we do is in large part determined by how we are, then in order to be responsible for what we do we have to be responsible for how we are. Robert Kane has argued that the lack of responsibility for how we are is the most significant lacuna in compatibilist accounts of responsibility. Kane sees the compatibilist as failing to take seriously, as Bernard Berofsky puts it, "our deep-seated yearning to be the ultimate source of our own natures."[200] Capturing this phenomenon, our sense of ourselves as ultimate sources, is what Kane calls the condition of Ultimate Responsibility (UR).[201]

But some compatibilists seem to worry about "sourcehood" conditions on responsibility as well. One way to understand Frankfurt's discussion of a hierarchy of values is as an outline of what conditions would have to be met to be responsible for one's values and therefore ultimately responsible for what one does.[202] It would seem, then, that UR is not a point of interest peculiar to any one side of the traditional debate. Nevertheless, incompatibilists like Kane are unlikely to be satisfied with a compatibilist answer to sourcehood concerns. They are likely to insist that compatibilist answers like Frankfurt's, while they might capture something of what is important about "owning" our inner selves, they fail to capture fully the notion of agents as "ultimate" sources. The disagreement, then, would hinge on how "ultimate" a source one must be of one's nature in order to be responsible. I focus here on Kane's stronger worry, since it more directly constitutes an objection to my account.

To have UR, then, requires being responsible for how one is mentally. But why should we require such a condition? The thought is that what we are responsible for doing can be explained by references to our beliefs and desires (and other mental states). These explain why we do what we do. But if these mental states weren't in some important way "up to us," then we would fail to exercise the appropriate control over them. They would seem to be "forced" upon us or "spawned" within us, and this would call into question our responsibility for them. But if they form the impetus and grounding

[200] Berofsky [2000], p.135.
[201] Kane [1996], [2005].
[202] See Frankfurt [1971]. My thanks to Manuel Vargas for directing me to this point.

for why we do the things we do, then this lack of responsibility for our thoughts seems to imply a lack of responsibility for what we do.

I think we are responsible for what we do (much of the time). Moreover, I think we are responsible for how we are (much of the time). Nevertheless, I don't think either of these claims depends upon our being the "ultimate sources" of our action, in the way UR implies. This section defends this claim. I begin by laying out Kane's view about UR and how agents might satisfy that condition. Then I show why even his argument fails to capture much of what is intuitive about UR (4.1). I then present Galen Strawson's argument for the impossibility of UR (4.2). Understanding this argument not only shows why "our deep-seated yearning" for UR may be just that, an unfounded desire, diagnosing why UR is impossible suggests a solution to the problem. But both Kane and Strawson seem to agree that a view without UR is in some way impoverished as a result. I conclude by showing why my view (and, in principle, any view) isn't worse for having failed to secure UR (4.3).

5.4.1. Kane, UR, and Self-Forming Actions

KANE FACES A PROBLEM. He believes that what we do is largely dependent on how we are, and he believes therefore that to be responsible for what we do requires being responsible for how we are. Furthermore, he thinks that in order to be responsible for how we are it would have to be the case that we somehow made ourselves that way. But this seems a difficult thing to do. As I noted when criticizing manipulation arguments, it seems that much of what we think comes to be largely without our influence, at least without our conscious control, in stark opposition to what we usually think is required for responsibility. Nevertheless, Kane believes that we do make ourselves to be certain ways whenever we perform what he calls a self-forming action (SFA).

A self-forming action occurs when an agent is forced with a dilemmatic-type choice, one between two alternatives of seemingly equal weight. Consider one of Kane's examples, that of a businesswoman who on her way to an important meeting sees a mugging taking place in an alley.[203] She experiences an inner struggle between stopping and calling for help (her moral conscience) and attending the meeting (her career ambitions). She resolves this struggle by turning back to help the victim.

Kane thinks that in such cases there is an indeterminate neurological struggle going on; what is essentially the physical manifestation of *two* separate psychological struggles. The first is to decide whether or not to attend the meeting, where the reasons for helping the victim act as "thwarters" to a solution; the second is to decide whether or not to help the victim, where the reasons for attending the meeting act as the "thwarters." Kane thinks that the businesswoman's own efforts of will to solve these two questions are the cause for the struggle. Her desire to advance her career prevents her desire to help from solving that decision, and vice versa. So, Kane thinks there's two processes going on, and that the results are indeterminate. Moreover, Kane thinks, when she turns back to help the victim, her choice, the "winning out" of her desire to help, means she is responsible for helping and for being the sort of person who helps.[204] In short, through self-forming actions an agent can come to be responsible for how they are with respect to

[203] Kane [1996], p.163-164.

[204] This is a simplification of the view. Kane actually believes it takes an unspecified number (but more than one) of SFAs before an agent has free will. But the picture above is intended to illustrate how it is that Kane thinks agents can come to be responsible for how they are.

the motive issuing in action. Helping the victim, on Kane's view, amounts to the businesswoman taking responsibility for her motive to help.

I must confess this strikes me as an odd account for the simple fact that it seems to mis-describe the phenomenology of such choices. The businesswoman isn't making two separate decisions; just the one, *between* helping *and* attending. Of course, the reasons for each have different sources, but that seems beside the point. Still, I'm willing to grant Kane his picture for my purposes here.[205] Instead, I want to criticize a different point. Let's suppose Kane is right, and individuals can come to be responsible for how they are through engaging in SFAs. The first remark is that it seems utterly plausible to think many people will not engage in any (or only a paltry few) SFAs in their lifetime.[206] An SFA requires an internal struggle, a conflict between two choices of equal weight. But we can easily imagine an individual who never faces such a conflict. She is moderately reflective, but always comes to reaffirm the beliefs and worldviews she grew up with. It's not as if she puts such views to tough scrutiny, though, she merely considers in various cases whether she ought to do differently than those views suggest and the answer is always a resounding "no." Indeed, her overall view is also consistent, so it isn't that circumstances are all that likely to produce a case in which her view itself contributes to conflict. Perhaps she grew up a utilitarian, and seeks always to maximize the good, and, by chance, has never encountered an instance in which she was pulled by competing forces. She has never engaged in sufficient SFAs in order to be responsible for how she is, but when she acts voluntarily, intentionally, and without mistake, she

[205] And, for all I know, the physical manifestation of that choice could fit Kane's description.

[206] Petit and van Inwagen have debated the rarity of SFAs. Here I obviously side with those thinking it likely many people would not experience enough to take responsibility for their mental lives on Kane's view. See Petit [2002] and van Inwagen [1989] for the respective positions. My thanks again to Manuel Vargas for pointing me to this debate.

surely is responsible for the outcome she brings about. So the first objection to Kane's picture is that only some people end up responsible for how they are. This is a rather unsatisfying picture. Indeed, I would hazard to suggest that quite a lot of individuals don't engage in SFAs in their lifetimes. This is obviously an empirical matter, but given Kane's characterization, I just don't think it all that likely.

My second observation is that Kane's view still doesn't secure what he was after. UR rests on an agent being the "ultimate source" of what he does, and this, it is thought (intuitively, at least) requires being responsible for the ways in which one is (mentally) that issue in action. But Kane's picture doesn't get us this. All he is able to secure is responsibility for some of the mental states of some agents. More specifically, he is able to show responsibility only for those mental states issuing in the outcomes of SFAs for those agents who engage in SFAs.

And even in those cases that fit Kane's model, it is hard to see how the sheer indeterminacy of the outcome is sufficient for rendering the individual responsible for the outcome. Consider Kane's businesswoman again. Suppose she was raised by divorced parents. Her mother always drilled a desire to compete and succeed into her. Indeed, the businesswoman's career ambitions are a direct result of her mother's influence. Her father, on the other hand, always sought to encourage his daughter to help others in whatever way she can whenever she can. Indeed, the businesswoman's desire to help the mugging victim is a direct result of her father's direction. Now, in the supposed case, the resolution of these two competing desires is indeterminate; we cannot settle ahead of time what she will do, which desire will "win out." But this fact alone doesn't seem to get Kane what he's after. For if the businesswoman isn't responsible for the competing desires, it is hard to see, to my mind, how the sheer indeterminacy of their conflict generates such responsibility by itself.

Kane may suggest that indeterminate neurological processes are the "source" we're after when looking to ground UR, that the control one must exert to be responsible for his mental states is just to have competing mental states conflict indeterminately. But it is hard to see why indeterminate neurological processes are "up to the agent" in the way UR was supposed to require. We also might notice that the businesswoman's motives in my elaborated example seem to have resulted in large measure due to her parents, and we might then ask why some indeterminacy in her brain makes her responsible for one of these motives. This result, to my mind, falls far short of the original goal. Perhaps Kane can still claim superiority over other views that don't get as near to UR, but this is a concessionary conclusion at best. But there is good reason for Kane's failure. UR is extremely difficult to attain. The next subsection illustrates just why that is.

5.4.2. Strawson on the Impossibility of UR

GALEN STRAWSON HAS given an argument on the impossibility of UR. His argument is helpful here for two reasons. First, it shows why Kane's project is likely doomed. Second, and more importantly, it helps draw out how the problem can be fixed. While we may not be able to secure UR of the sort Kane seeks, we can establish a weaker sort of responsibility for how we are, and this may be all that we need. I'll consider this last point in the final subsection.

Strawson's Basic Argument is as follows:[207]

1) You do what you do because of how you are (mentally).

2) If you do what you do because of how you are, then to be *ultimately* responsible[208] for what you do requires being ultimately responsible for how you are.

But,

3) You cannot be ultimately responsible for how you are.

So,

4) You cannot be ultimately responsible for what you do.

This is the Basic Argument in its simplest form. In defense of Premise 3, Strawson offers the following:

5) If you are ultimately responsible for how you are (call this psychological set X), it must be because you intentionally brought it about that you had X.

6) To have intentionally brought it about that you had X, however, requires a previous way that you were (call this psychological set Y, including, at least, an intention to make yourself have X).

But,

[207] Strawson gives four separate versions of the "Basic Argument", which he thinks are all restatements of each other, some with more detail than others. I have summarized them into the version I present above, which I think effectively captures the main thrust of Strawson's argument, and makes clear the place of the argument in my overall discussion. For his four versions, see G. Strawson [2001], pp.443-449.

[208] Strawson actually uses the term "URD", which means "ultimately truly and without qualification morally responsible and deserving of praise or blame or punishment or reward," where I have used the simplified UR (meaning "ultimately responsible"). I will discuss the differences between these two uses in Section 4.3. My simpler version is sufficient, I think, and does no disservice to Strawson's argument.

7) If you are ultimately responsible for having Y, it must be because you intentionally brought it about that you have Y.

And,

8) To have brought it about that you had Y requires a previous way that you were (i.e., psychological set Z).

9) And so on…

Strawson believes this to be an infinite regress that shows ultimate responsibility for how we are to be impossible, and thus that ultimate responsibility for what we do is also impossible, since the latter depends on the former.

Premise 2 captures what UR is all about; a yearning to be the "ultimate source" of our actions. To be truly and deeply and most meaningfully responsible for what we do, we'd have to be responsible for how we are in a true, deep, and meaningful way. The argument essentially relies on the conflict between two premises to get the regress going. Premise 2 sets out the conditional that demands responsibility for how one is, and Premise 5, which claims that responsibility for how one is requires some action on the agent's part. Thus, Kane thinks SFAs are something an agent can do to become responsible for at least some aspect of the way they are.

Interestingly, Strawson does not argue for either Premise 2 or Premise 5. He takes them to either be obvious or self-evident upon reflection.[209] For the moment, I'll grant Strawson these assumptions. Notice that this is all it takes to generate the regress. If we have to do something to be a certain way, and to bear a certain relation to what we do requires bearing that relation to the way we are, the regress immediately follows. For every way we might be that could bear the relation, we'll always have to look for what

[209] See G. Strawson [2001], p.443-446.

we did to get that way, which in turn will suggest a new way we are. To specify this a bit: for each motive we might be responsible for, we'll always have to look for what we did to come to be responsible for it, and that action will suggest a new motive on which we acted. The regress follows.

There are two interesting conclusions to take away from this. The first concerns the prospects for UR. Kane suggests that we have a "deep-seated yearning" to be the "ultimate source" of what we do. This, he and others think, requires responsibility for how we are. Strawson has extended that notion of deep responsibility to the way we are and shown that it would be impossible to achieve it. But notice that it seems that not only could humans as we know them never be achieve UR, few beings could. At least, it would be a mysterious sort of power to be able to manage UR, for it would require being responsible for how you are from the moment you are that way, and presumably from the moment of existence. This looks more like a supernatural power than the sort of ability ordinary creatures might have. Of course, this isn't a knockdown argument against UR; perhaps agents do have special abilities.[210] The conclusion to draw, I think, is that UR isn't the sort of responsibility we should be looking for. Indeed, our deep-seated yearning may be just that, an unfounded desire to be some sort of *causa prima* of ourselves (or at least a *causa sui*).[211] This standard is plausibly too high for most beings, much less imperfect creatures such as ourselves, ones who often come to have beliefs and desires and thoughts quite inadvertently (or without conscious control). Obviously, this is Strawson's point: it would be impossible for us to be this way. But far from leading us

[210] The literature on agent causation claims that agents are capable of a special sort of causation, one that grounds free will and responsibility. For a summary of such views, see Clarke [2004], Section 3.

[211] One of the versions of Strawson's Basic Argument turns on the notion of ultimate responsibility requiring us to be *causa sui* (meaning "the cause of itself").

to the conclusion that UR is responsible, I should think it invites a certain skepticism that he’s described a notion of responsibility worth wanting.

The second interesting conclusion to take away concerns the prospects of rejecting of the regress’ two essential premises. If we reject either that responsibility for what we do requires responsibility for how we are or that to be responsible for how we are requires that we do something to make ourselves that way, then we can avoid the regress. We may not end up with UR, we won’t be the “ultimate source” of what we do, but that doesn’t mean such a view is worse off for it. I think Premise 2 is likely flawed. But I will not argue against it here.[212] I only need to present the case for rejecting one of the two premises, and I believe we already have some reason for rejecting Premise 5.

Premise 5 requires that we do something consciously to bring it about that we are a particular way in order to be responsible for being a particular way. This is a natural thought. The paradigm for responsibility is voluntary, intentional action. So if we want the conditions on responsibility for how we are, why not just use the conditions on responsibility for what we do? But there is a glaring asymmetry between these two phenomena, one we already uncovered in discussing manipulation arguments. It doesn’t seem as if many of the thoughts we have we bring about advertently. I’ve suggested that we virtually never have robust control over forming a belief, and I think this holds true of many of our other mental states. But if this is true, it seems foolish to require that intentionality is necessary for responsibility for those states. It would be a fool’s errand, as Strawson’s Basic Argument so starkly illustrates. Instead of simply using our conditions on action for mental states, the more promising strategy would be to use our methodology for arriving at those conditions on action in the first place. Recall that we

[212] For a critique of both premises, see Clarke [2005].

looked to those considerations that seem to undermine responsibility (or blameworthiness and praiseworthiness). Then we thought that perhaps each type of consideration suggests the absence a necessary condition on responsibility. So to get the conditions on responsibility for what we do, we needed to look at the undermining factors in cases of action, and work our way back to conditions those considerations suggested were absent.

A similar methodology might work in the case of responsibility for how one is. The first thing we would note is that advertence doesn't seem as if it would be required, for the very reasons we've already discussed. Suppose Clyde is a bigot. He hates minorities. And suppose that he came to have this belief from his upbringing, coupled with a few bad experiences in his early teens. A few times he's sat down and reflected on whether he should be prejudiced and decided that the evidence is in favor of being that way. Now, it seems to me that Clyde is surely blameworthy for his bigoted attitudes. But suppose Clyde has been reading up on his philosophy. He knows of Strawson's Basic Argument, and of the points I've made about manipulation arguments. Clyde reasons that his coming to his racist attitudes weren't really under his control. At every step his was a captive to the way in which the situation struck him, and though he took steps to consider how he ought to be, the answer never seemed in doubt. Still, Clyde is blameworthy for his attitudes. His appeals to the inadvertence of his attitude acquisition stand as no defense. And this suggests that advertence is not required.

What is required? I'm not prepared to provide an extensive defense of any view here. I'm not particularly committed to one. My preferred method would be to examine the considerations that would undermine our ascriptions of responsibility for mental states. Such a task would be too lengthy for present purposes, however. Instead, I'll offer a preliminary response, one that I think is natural, though I won't explicitly defend it. We can get a sense of the sort of view that might work by holding some sort of identification view. Roughly, suppose A is responsible for attitude x so long as A identifies with x in some way suitably specified. Obviously, *much* more would have to be said about what identifying amounts to, but so long as it wasn't an action (and I see no reason to suppose the notion would have to be an action), then the regress no longer threatens. Plus, I suspect that failure to identify with an attitude could be one of the few considerations that might actually undermine responsibility for that attitude. Much more needs to be said in defense of such a view; I merely sketch it here in the hopes of suggesting a plausible way out of the regress.

Now, Strawson might suggest that inadvertence *should* undermine responsibility for mental states. He might insist that responsibility for how we are should be the same sort of notion as responsibility for what we do, and we should expect the same conditions to govern both. Thus, my examples of inadvertent belief only serve to further undercut confidence in responsibility. I think the notion of responsibility for how we are is similar as the notion responsibility for what we do, but I think that they concern two radically different classes of objects: mental states on the one hand, actions and outcomes on the other. Moreover, Strawson's argument concerns only UR, not responsibility in any reasonable form, that is, of a sort that might serve a suitably explanatory theory. If a theory can preserve a notion of responsibility that explains the core phenomena we're concerned with, it is a small criticism that it doesn't explain everything one might be interested in. So, evaluation of this concern would then rest on evaluating whether UR represents a core phenomenon worth explaining, or perhaps is not so central after all.

The worry remaining, then, is that in avoiding the regress we've been forced to weaken our notion of responsibility such that we no longer get UR. If so, Kane may be correct in thinking that a view such as mine fails to capture an important and compelling phenomenon of our reflections about responsibility. Section 4.3 addresses and rejects the concern that a view such as mine is impoverished as a result.

5.4.3. Ultimate Responsibility vs. Responsibility

IN GENERAL, I THINK, compatibilists are likely to be confused about the desire to be the "ultimate sources" of our actions. Once we fill out a bit what it would take to be an ultimate source, it looks like perhaps such a notion requires the sort of "panicky metaphysics" compatibilists are traditionally wary of, in part because it suggests that agents have mysterious powers. If so, it's hard to take seriously failing to win UR. But this isn't a victory for incompatibilists or impossibilists[213] if it turns out UR isn't worth having. This would suggest that it is the wrong notion to be after in the first place, rather than highlighting a deficiency of a particular type of view.

It is all too likely that much of what determines how we are is beyond our control. Our genes and other biological factors no doubt contribute, as do social and environmental factors, especially as we develop into adults. So it should be no surprise that we would have difficulty in establishing ourselves as the "ultimate sources" of our actions. But we can still be the sources of our actions, and we can still be responsible for how we are and for what we do, even if it isn't for the same reasons in both cases.

[213] Impossibilists hold that responsibility (specified in some way) is impossible. One can be an impossibililst for a variety of reasons, but these normally do not turn on the thesis of determinism, since an impossibilist would hold that responsibility is impossible in all possible worlds, even those that are indeterminate.

Moreover, I think that a picture of responsibility for how we are that rests on identification (of some sort or other) with our attitudes provides a natural and plausible story, one that meshes well with a story about responsibility for outcomes. Despite the influence of genetic and environmental factors, most ordinary agents tend to come to identify with their psychological sets. And in the same way that adults come to be responsible for more of what they do, so too do they become responsible for more of how they are. We may like to change some tendencies (like being forgetful or being quick to anger), but we also often accept these tendencies as part of who we are.[214] We also tend to be proud of certain attitudes and dispositions (e.g., kindness or adventurousness). I think a theory is impoverished if it cannot account for these phenomena; that we can justifiably be proud of certain traits, and it is no mistake to see others as blameworthy for them. I think this requires responsibility for how we are. But it doesn't require UR.

The previous section illustrated that the insistence that the conditions on how we are match those on what we do leads to a regress. We can avoid the regress by giving different conditions on responsibility for how we are, and I've sketched a rudimentary position concerning an agent's identification with his attitudes. Such a view I think fits nicely with the conditions on responsibility for outcomes, and a natural story about how responsibility "increases" with maturity. Though many more details need to be developed, there is, I think, a promising position here.[215] Perhaps we would be better off if we could secure UR; it does seem that having such abilities would make us more powerful. But it isn't clear that failure to secure such powers impoverishes a theory. Instead, I think separate stories about responsibility for outcomes and responsibility for

[214] We are especially apt to do so in relation to those we care about, forgiving them their flaws, or even loving them flaws and all.

[215] Working out the details is a project that I'd like to work on after the dissertation.

how one is are sufficient for our purposes. We can explain all the relevant phenomena save our "deep-seated yearning" for UR. If we instead focus on ordinary agents, a plausible and natural picture of responsibility emerges, one that seems sufficient for creatures like us.

5.5. Alternate Possibilities Condition

THERE IS A FINAL CONDITION waiting in the wings, the veritable 800-lb. gorilla of proposed conditions necessary for responsibility. Many have claimed that if the agent could not have done otherwise, then he isn't responsible. Therefore, responsibility requires that the agent have genuinely alternate possibilities (or, the ability to do otherwise).[216] The literature (both for and against) concering this condition is vast, and full discussion of it would require much more attention than can be given it here. Authors like van Inwagen have urged that responsibility requires an ability to make some fact not be the case. To be able to do otherwise then break a vase, I must have the power to make it such that the vase doesn't break (under the present circumstances). Some, most notably Frankfurt, have argued that agents can be responsible even if they couldn't have done otherwise than they did. And some authors, most notably Fischer & Ravizza, hold that even though one can be responsible without alternate possibilities, one cannot have free will. Sorting through these various approaches, the slightly different construal

[216] Historically, classical compatibilists have argued that an agent could have done otherwise if he *would* have done otherwise had he *wanted* to do so. This is known as the 'conditional analysis.' Without digressing to the debate over whether such an analysis is legitimate or useful for moral responsibility, I'll simply specify here that the meaning of the phrase I wish to consider is that endorsed by incompatibilists, who wish it to mean that given the same initial conditions (including an agent's desires, e.g.) it is truly indeterministic (or genuinely open to the agent) what the agent will do. For a good survey of relevant articles and issues, see Widerker and McKenna [2003].

each gives to the requisite condition, and the litany of objections and replies, would far outstrip our discussion here.

Instead, I propose to show (1) that the Alternate Possibilities condition is undermotivated, and (2) that the core of Frankfurt's original insight is true. The conjunction of these two claims is sufficient, I think, to place the burden on the defender of the AP Condition, rather than on a theory of responsibility that lacks such a condition.

5.5.1. What the Condition Requires

TO GENERALIZE, THE ALTERNATE Possibilities Condition states that to be responsible for some action or outcome, the agent must have had genuine alternate possibilities open to him at the time of action. The condition derives its plausibility, I think, from an intuitive self-conception of choice. It seems to us that when we deliberate we face a "garden of forking of paths,"[217] a series of genuine choices, where it is open to us to go left or right, say, at each fork. If we were in a maze of hallways and doors, where at each choice one door was always locked, it seems as if we'd have no genuine choice at all. Indeed, we would be "forced" to take the unlocked doors at each junction, and the result, whatever it was, would not be of our own making. It would be determined by the arrangement of the locked doors. This picture of deliberation and choice, it seems, supports a condition that requires a genuine possibility of choosing differently than one does. To be responsible for taking Door #1, I must have been able to take Door #2. In short, I had to have been able to do otherwise.

[217] This is a phrase introduced by John Fischer, borrowing, I believe, from Jorge Luis Borges.

If the above picture is right, then my account would require modification. I would have to add the additional condition that the agent could have done otherwise than he did. But I don't think the above is right. I think the Alternate Possibilities Condition is undermotivated. That is, we need to have good reasons for thinking that agents without genuine alternate possibilities are not responsible *because* they lack alternate possibilities. But, I argue, we have not been given these reasons yet. So, until it can be shown that the lack of possibilities itself undermines responsibility, we have insufficient evidence in support of the Alternate Possibilities Condition (AP Condition), despite its pretheoretical appeal.

My argument is thus a burden-shifting argument. I don't think I'll convince those who find the AP Condition deeply intuitive, because it is rather hard to shake loose of such judgments. Nevertheless, in the larger dialectic surrounding the significance of the AP Condition, we always face conflicting judgments. Incompatilibilists are more likely to think scenarios without genuine alternate possibilities lack responsibility, while compatibilists are much more likely to find responsibility preserved. So we need something more than pretheoretical intuitive plausibility to tip the balance. My argument here, therefore, doesn't show that responsibility doesn't require genuine alternate possibilities. It is rather intended to show that the case for the AP Condition rests on pretheoretical conceptions of human action, not on careful consideration of core cases of undermined responsibility. While such cases are not the only relevant data points, we have already seen how fruitful their examination can be, and I again turn to these cases to show that the AP Condition is undermotivated. If I'm right, then an account that explains responsibility without such a condition is not obviously lacking. The burden then rests on the defender of the AP Condition to show that its absence is damning for any theory of responsibility.

5.5.2. Examining the Evidence for the AP Condition

WHY MIGHT WE THINK that responsibility requires genuine alternate possibilities? One reason one might think so is by citing cases in which the agent doesn't seem to have genuine alternate possibilities and isn't responsible *for that reason*.[218] But such cases, I think, are often misrepresented. I'll consider a few such cases here.[219]

For instance, recall my discussion of coercion. There, I noted that in cases of coercion people often say that they didn't have a choice. They couldn't have done other than they did on pain of death. To repeat my claim there, however: this is really hyperbole. It isn't that coerced agents don't have a choice; it's that they don't have a *good* choice. The alternatives open to them are all bad in one way or another. So it is a mistake to think that coercion implies the lack of genuine alternate possibilities. Even if we concluded (contra my argument from Chapter 3) that coercion does undermine responsibility, it wouldn't be because it eliminates genuine alternate possibilities. It would have to be for some other reason. Therefore, cases of coercion do not support the AP Condition.

We might instead look to cases where the agent is incapacitated. For instance, suppose Irene is tied up in her basement. She promised to meet her friend for lunch, but

[218] A similar observation is central to Frankfurt's original article, Frankfurt [1969], pp.18-20.

[219] It is remarkable, I think, that so little attention has been given motivating the claim that responsibility requires the ability to do otherwise. It is often assumed to be obvious and uncontroversial. Indeed, even those who critique Frankfurt's argument still take it as intuitive that responsibility requires alternate possibilities.

fails to, since she cannot escape her bindings. Here we may be tempted again to claim it is the lack of genuine alternate possibilities that undermine Irene's responsibility for breaking her date. We might think that Irene couldn't but fail to meet her friend, since she was unable to do anything but remain tied up. But it isn't obvious that it is due to this inability that Irene's responsibility is undermined. If we have plausible alternative explanations, for example, that Irene doesn't choose to miss the date or that her breaking the promise evinces no ill will, then pressure is placed on the defender of the AP Condition to explain just why it is Irene's lack of genuine possibilities that does the work.

Still another set of possible exemplary cases concern compulsion. If the urges compulsions like kleptomania or agoraphobia present are really irresistible, perhaps our tendency to excuse, say, kleptomaniacs counts as support for the AP Condition. In Chapter 4, I suggested that our reasons for not blaming kleptomaniacs may come from the fact that we ought not blame them rather than with facts concerning their responsibility. Nevertheless, I didn't assume there that compulsions really were irresistible. How likely is this claim? First, it is important to note that a kleptomaniac will not always steal no matter the situation, nor will an agoraphobe remain inside regardless of his circumstances. If there is a security guard standing right in front of the kleptomaniac, then perhaps he won't steal. Or, to take a bizarrely contrived example, suppose that if he eats waffles for breakfast, he is able to resist his urges to steal. Similarly, we can suppose that there are some particularly strong reasons that might get an agoraphobe out of the house; for instance, setting his house on fire.[220] We should conclude, therefore, that the mere presence of a pathological compulsion is not by itself enough to show that the agent could not do otherwise.

[220] This example is borrowed from Al Mele.

It would seem, then, that ordinary instances of potentially undermining factors do not provide evidence for the AP Condition. They do not suggest that the agent's responsibility is undermined due to the lack of genuine alternate possibilities. It is not from examining these core examples that leads one to suspect that responsible agents must be able to do otherwise. Instead, it is more likely that the AP Condition is supported mainly from our pretheoretical conception of deliberation and action. It *seems* to us that our futures are "gardens of forking paths," and that a vital power of human agency is the ability to choose between genuine alternatives.

Suppose most people do believe that genuine alternative possibilities are necessary for responsibility. It is highly likely that this is because they both believe that they in fact have genuine alternate possibilities and that most agents are responsible for most of what they do. That is, it is *because* they are antecedently committed to responsibility and the ability to do otherwise that implies a commitment to the AP Condition. Suppose also, as is often the case, that the truth of determinism would mean that we do *not* have genuine alternate possibilities.[221] Now, we have several possibilities. It could be that most people do not believe human action is determined. This is a natural assumption, especially given their commitment to the AP Condition. We might then wonder what would happen if most people were convinced of the truth of determinism? It seems they could either give up the AP Condition or give up responsibility. One of Strawson's key insights was that people are much more likely to retain a commitment to responsibility itself than to any conditions on it. Indeed, without independent support for the AP Condition, its intuitive appeal may simply rest on a mistaken picture of human

[221] This is, as with most claims about the ability to do otherwise, a controversial claim. But it conforms to the typical layout of the debate, with incompatibilists most in favor of the AP Condition, and using it to argue that determinism conflicts with responsibility as a result.

action. And we may welcome a theory of responsibility that doesn't require so strict a requirement for its obtaining.

5.5.3. Frankfurt's Key Insight

FRANKFURT HAS MOST famously contested the AP Condition. His examples take on a simple structure and have come to be known as Frankfurt-type cases.[222] We can construct such a case by returning to our previous example involving the maze of doors. Suppose that at each junction, there are two doors, one of which is always locked. Thus, for any choice between doors, an agent can only go through one; the locked door prevents him from actually choosing that door (or what lies behind it). It would seem, then, that an agent in such a maze lacks genuine alternate possibilities. At every junction, there is only one door he can open. Now, suppose an agent goes through the maze always selecting unlocked doors. At every junction, he deliberates, and chooses the unlocked door. It is of course somewhat lucky for him, since he only ever has a 50% chance of choosing "correctly." Despite his odds, however, he always chooses the unlocked door. Frankfurt asks us to consider why such an agent isn't responsible for his path. He deliberates and chooses to go through a particular door. Of course, he had to go through that door, since the other door was locked, but why should these counterfactual facts matter, when in the actual sequence of events[223] he chose the door he did.

Frankfurt's key insight, therefore, is that when considering ascriptions of responsibility, it is sufficient to examine what was true of the agent in the actual sequence of events, whatever counterfactual facts might hold. Our agent in the maze lacks genuine

[222] John Locke many, many years before Frankfurt identified the same insight, but did not develop it as Frankfurt has.

[223] To my knowledge, Fischer and Ravizza are the first to make explicit the distinction between counterfactual considerations and the "actual sequence."

alternate possibilities, but when he chooses to take a particular open door, the fact that the other door was locked does not obscure his responsibility. We can make the example more palpable. Suppose Jones wants to injure Smith.[224] So he plans how to get Smith alone in a secluded spot, does so, and then beats him severely. Now suppose that Black also wants Jones to injure Smith, and has a device that allows him to exercise some control over Jones' brain. Should Jones fail to decide to attack Jones, then Black will activate the device, making Jones decide to attack Smith. As it happens, however, Jones deliberates and decides to attack Smith, and Black never needs to use the device. Again, as in the maze case, it seems that no matter what Jones does, he will choose to attack Smith. His action is unavoidable. Yet, despite this fact, in the actual sequence of events, Jones deliberates and chooses to attack Smith. Frankfurt thinks that in such cases we cannot find fault with an ascription of responsibility. Jones is surely responsible and blameworthy for injuring Smith, even though he could not have done otherwise.

Frankfurt's conclusion has been contested. But I need not endorse his conclusion to make my point. Frankfurt's key insight is that we don't examine the counterfactual scenarios when evaluating an agent's conduct. Jones kills Smith deliberately and purposefully. Nothing else is required for an ascription of responsibility, and indeed, blameworthiness. Had Black intervened, or had our agent in the maze tried a different door, then perhaps our conclusions would change. But given the actual scenarios, we have sufficient cause for ascribing responsibility.

[224] This formulation of a Frankfurt-case is more traditional, since it involves a "counterfactual intervener" (Black), someone who will intervene to ensure something happens, but in the actual sequence doesn't need to. I prefer cases like the maze example above because Black is an Immune Intervening Agent, and can adversely affect interpretations of the case for the same reasons as in manipulation arguments. Those problems can be safely ignored in the present context, however, and this traditional example gets our moral engine running since it is a morally loaded case, as opposed to the opening of doors.

Responsibility may still require the AP Condition. But the case must be made for this condition, and it must rest on more than a pretheoretical notion of human action. For we may well be wrong about such a picture. And given theoretical reasons for examining core cases of undermined responsibility, and the fact that such cases do not imply that the lack of genuine alternate possibilities undermines responsibility, the case of the AP Condition appears undermotivated. When combined with the crux of Frankfurt's insight, I think there is sufficient pressure to place the burden with the defender of the AP Condition to show why responsibility requires the ability to do otherwise, rather than with theories that lack such a condition. I take it, then, that if my account succeeds in all other respects, the AP Condition will not stand as an objection itself.

Chapter 6: The Capacities of Ordinary Agents

6.1. Introduction

IN CHAPTER 1, I outlined what I take to be the Strawsonian Compatibilist Program, which is a two-pronged endeavor.[225] The First Prong is to explain the metaphysically worrisome notion of responsibility in terms with which we are much more familiar, namely, blameworthiness and praiseworthiness. The program understands these notions through our normative practices of blaming and praising. The Second Prong is to distill the conditions on responsibility from those considerations that undermine it. So long as no such considerations are present, the agent is responsible. Part of the aim of Chapter 2 was to show that we have good reason to reject the First Prong because of careful examination of the Second Prong. Once we attend to the considerations that undermine blameworthiness and praiseworthiness, we see that the same considerations undermine

[225] Both of these prongs owe their origins to Strawson's original essay, "Freedom and Resentment." On the First Prong he says, "I want to speak…of something else: of the non-detached attitudes and reactions…of offended parties and beneficiaries; of such things as gratitude, resentment, forgiveness, love, and hurt feelings" (p.62). On the Second Prong he says, "Then let us consider what sorts of special considerations might be expected to modify or mollify this feeling [i.e., the reactive attitude of resentment] or remove it altogether" (p.64).

both notions symmetrically. Indeed, the same considerations appear to operate symmetrically across an immense range of cases, even in non-moral cases of blameworthiness and praiseworthiness. The hypothesis I drew from the data was that blameworthiness and praiseworthiness share some important component that the undermining factors operate on. And since it seems an individual must be responsible for something in order to be either blameworthy or praiseworthy for it, I concluded that in order to explain the wide set of undermining cases an account must appeal to some notion of responsibility that is both explanatorily prior to blameworthiness and praiseworthiness and significantly independent of our practices of praising and blaming.

The remainder of the dissertation has been a development of the Strawsonian Program's Second Prong. Rather than settle for a negative condition on responsibility (that when no undermining factor applies, the agent is responsible), in Chapter 3 I suggested three positive conditions on responsibility. If we closely attend to a categorization of the undermining factors, sorted by the specific features of agents and actions that seem to be missing in cases of undermined responsibility, we can posit the presence of those same features as being necessary for responsibility. Still, I retain the basic negative notion of Strawson's Second Prong by insisting that so long as all three conditions are met, we need not require any more for responsibility. The subsequent two chapters were a defense of these conditions as individually necessary and jointly sufficient. Chapter 4 defended them as necessary for responsibility from the objection that negligent agents fail my conditions but are still responsible for outcomes. And Chapter 5 defended them as jointly sufficient from a set of objections suggesting that my conditions left something out.

This chapter has a very different aim. In the long-standing debate between compatibilists (who think responsibility is compatible with the truth of determinism[226]) and incompatibilists (who think it is not), it is of crucial importance what capacities humans must possess in order to be responsible. For if individuals don't have the capacity to fulfill the conditions on being responsible for things, then it would seem that we could never be responsible for anything. Thus, hard incompatibilists (incompatibilists who believe we are never responsible) think that whatever the conditions on responsibility are, the world is such that we never satisfy them. One standard way of demonstrating this is by illustrating that, were determinism true, we wouldn't have the capacities needed to satisfy the conditions. For example, if one thinks the ability to do otherwise is required for responsibility, then one could attempt to show that a deterministic universe would make such an ability impossible.[227] An immense amount of ink has been spilled over whether or not we have such metaphysical capacities as being able to do other than we did.[228]

Within the dialectic, then, there is a premium on the demandingness of the conditions on responsibility. What must the world be like for creatures such as we are to be able to satisfy the conditions on responsibility for particular things? Are they such that the truth of determinism would put responsibility beyond our reach? Compatibilists are committed to answering "no;" the world as it's currently configured is such that creatures like us are responsible most of the time (or at least could be). Or, to put it

[226] Determinism itself is the thesis that all of the present facts about the world plus the laws of universe entail all future truths (including truths about human actions and choices).

[227] Van Inwagen, among others, gives an argument like this. See van Inwagen [1975]. For a similar view I've already discussed (but doesn't rely on determinism being true), see G. Strawson [2001].

[228] To get an idea of the extent of this debate, one needs only look for the articles spawned by Frankfurt's original paper. See Frankfurt [1969; 2003], for his original statement, and Widerker and McKenna's [2003] for a sampling of the secondary literature.

another way, it must be the case that even in the face of determinism, creatures like us have the necessary capacities to meet the conditions on responsibility, whatever these happen to be. It is to the compatibilist's advantage, therefore, to have conditions that are easy to meet. The easier they are to meet, the less likely determinism will pose a threat to their satisfaction.

Part of what has been so appealing about the Strawsonian Program is its ability to simplify the conditions on responsibility. By avoiding a metaphysical notion of responsibility, the Strawsonian Compatibilist avoids much of the intractable debate surrounding whether or not agents have those metaphysical abilities such a notion requires. Whether or not agents actually have, say, the robust ability to do otherwise, it is surely no uncontroversial matter. Indeed, no matter which position one takes, all parties would seem to agree that demonstrating the truth or falsity of such a claim is no easy task. So it has been a virtue of Strawsonian Compatibilism that it side-steps this thorny issue, by mediating explanation of responsibility via our social practices and the norms governing them. Strawson originally put the point in terms of our confidence in holding each other responsible. Given the norms governing our practices, we should be confident that we're responsible, and this confidence would remain unshaken even were we to learn of determinism's truth.[229] Thus, on Strawson's original proposal, all we need to be responsible is not to be subject to any undermining factor. If the outcome wasn't an accident, or inadvertent, or a mistake, etc., then the agent is responsible. Determinism has no grip here. These conditions are quite easily met, and don't appear to rest on any "panicky" metaphysics.[230]

[229] See Strawson [1962], pp. 73-76.
[230] To use Strawson's phrase. Strawson [1962], p.80.

My compatibilist view rejects the Strawsonian Program's First Prong, the one that avoids the metaphysical worries. I claim that we need an independent notion of responsibility, one not mediated through our practices. And, indeed, there are positive conditions that must be met in order to be responsible for outcomes on my view, so it isn't just the case that we can perform a check to see whether or not any of the undermining factors apply. One might fear that as a result my compatibilism gets us a better account of the undermining factors only to lose compatibilism to the dangers of the metaphysical debate. Again, whether or not it can be demonstrated that we have the requisite abilities, a theory that can do the work without taking on such a difficult task would seem to have quite attractive advantages. Thus, if my brand of compatibilism is to be a serious alternative, I must find a way to retain this attractive quality of the Strawsonian Program. I must show that our confidence that we are responsible for much of what we do is well-founded; that metaphysical worries about our abilities won't get enough traction to prove problematic even in the face of determinism.

Fortunately, we should be confident that we are responsible for much of what we do. My three conditions are quite easily met. Indeed, they are met so long as human agents possess three corresponding capacities (which I'll outline below in Section 2). So, we should be confident that we are responsible to the extent that we are confident that we are human agents possessing those basic capacities. Thus, while my view rejects the Strawsonian Program's particular method for avoiding the metaphysical entanglements associated with free will and responsibility, it retains the basic attractive advantage of the program's position: we should be highly confident that we are responsible for much of what we do.

6.2. Conditions and Capacities

I HAVE STATED THAT an agent is responsible for an outcome just in case: 1) the outcome was brought about voluntarily; 2) the outcome was brought about intentionally; 3) the outcome was brought about without mistake. These conditions are really quite weak. They are so weak, in fact, that all ordinary agents satisfy them most of the time. The Voluntariness Condition can be met so long as most actions of agents are explainable by belief-desire sets. The Intentionality Condition can be met so long as agents can foresee the consequences of their actions (or that they recognize they can affect the world so as to bring specific things about). And the No-Mistake Condition can be met so long as agents have the capacity to recognize factual features of their actions. So, for it to be the case that ordinary agents *can* be responsible, ordinary agents must possess the following three capacities: (a) a capacity to act on belief-desire sets; (b) a capacity to foresee consequences of their actions; (c) a capacity to recognize factual (non-moral) features of their actions. These capacities need not be constitutive of agency, nor exhaustive of it; all I require is that, whatever else is true of them, ordinary agents indeed possess these three capacities. Moreover, these capacities need not be infallible, or always engaged, or applied in every instance. Indeed, the capacities presumably won't all be exercised precisely in cases of undermined responsibility. All that has to be true is that ordinary agents have such capacities and that they exercise them regularly. If this is true, then I think it clear that ordinary agents can be responsible for much of what they do.

How plausible is it, then, to suppose that ordinary agents possess all three capacities? I find it eminently plausible; indeed, it seems simply obvious that we have such capacities, and that we exercise them regularly. Not to belabor the point, but I do want to emphasize just how uncontroversial these capacities should be. They should be uncontroversial largely because to imagine a creature who lacked any one of them is to imagine a human far outside the norm. The capacity to act on belief-desire sets is satisfied so long as belief-desire sets function in explaining action. Such a capacity need not specify how beliefs and desires function in the actual etiology of action, beyond committing itself to some causal role. This capacity, then, is compatible with a large range of realist theories about the mind.[231] Furthermore, someone who lacked this capacity would be a strange creature. Even lots of animal behavior can be explained by reference to their belief-desire sets, so a human being who lacked such a capacity would be an aberration indeed.

The capacity to foresee the consequences of one's actions is just the capacity to recognize the differences one makes (and can make) to the world. We recognize that we can affect the world in certain ways, either by taking steps in order to fulfill our ends or as consequences of our taking steps to fulfill our ends. My dog may not realize that in jumping on the couch to bark through the window at the mailman he's damaging the fabric, say. Indeed, my dog may not have the capacity to recognize that in fulfilling certain desires, like barking at the mailman (or, say, defending his territory), he can affect the world in other ways (like damaging the couch). Unlike our dogs,[232] however, we do

[231] Someone like Dennett, who takes an interpretationist stance towards beliefs and desires, might find this commitment contentious. On an interpretationist view, the capacity above could be modified so as to merely require that the belief-desire set be explanatorily efficacious, not that it actually cause behavior. If Dennett turns out to be right, then the following worries I discuss won't turn out to be worries at all. But since I think that this view of the mind is more controversial than realism about beliefs and desires, I've opted for a commitment to beliefs and desires being real and playing a causal role in action.

[232] This point is speculative. I use dogs as an illustrative example, rather than a substantive one. I intend it only to show the distinction. This caveat notwithstanding, I'll assume that at least some non-human animals lack this capacity.

have this capacity. It’s the one that let’s us plan in complex ways for complicated and far-reaching goals. And it’s the self-awareness we have, that when we do absent-mindedly or by accident cause harm or bring about some effect, we recognize that it is our doing. Now, possession of such a capacity doesn’t amount to satisfying the Intentionality Condition on responsibility; that requires a conscious entertaining of the possible outcomes. But regular satisfaction of that condition seems quite reasonable given this general capacity and our experience with human action. Note as well that someone who lacked such a capacity would not see herself as having an effect on the world around her. It seems difficult to even imagine this, and yet, if it were to happen, it would surely seem to evince a terrible cognitive deficiency in such a subject. Again, such a creature would be a severe aberration from the norm.

Finally, the capacity to recognize factual features of their actions is just a capacity to possess appropriate concepts relevant to one’s action. Those who possess such a capacity are able to see that they are taking someone else’s property, or that pushing the button will cause pain, or that driving this fast is dangerous. Such a capacity need not be infallible; we can, of course, make mistakes. And we can fail to consider factual features of our action, as in cases of inadvertence. But the general capacity remains nonetheless, just as I retain the capacity to ride a bike though I’m presently too dizzy to succeed at this moment. One who lacked such a capacity might be able to act, and might see herself as affecting the world, but would fail to see the different ways she was affecting the world. In short, it would seem she’d be unable to see how she fit into the world, and how, in turn, the world itself fit together, at least from the human (agent?) point of view. Again, I take it as painfully obvious that such a creature would be a radical aberration from ordinary agents.

If these three capacities really are quite uncontroversial, then it should be apparent that satisfaction of my three conditions on responsibility, which rest on these capacities, should be easy to come by. We should have extraordinary confidence, therefore, that ordinary agents, since they do possess these capacities and exercise them regularly, can be responsible for most of what they do (naturally, barring instances of the undermining factors). And we should therefore only be skeptical about responsibility for outcomes to the extent that we're doubtful such creatures as these ordinary agents exist.

6.3. Future Threats

IT SHOULD EQUALLY BE obvious that the truth of determinism poses no special threat to my conditions on responsibility. Even were our actions determined by past facts and the laws of nature, the general capacities underwriting the conditions on responsibility would be unharmed. No feature of our actions being determined would suggest that we don't act on belief-desire sets, or that we can't foresee the outcomes we bring about, or that we can't have action-relevant concepts. Thus, in rejecting the Strawsonian Program's First Prong we do not commit ourselves to a more demanding notion of responsibility, and we therefore do not lose its attractive advantages. My compatibilist conditions seem easy to meet and determinism poses no special threat to their satisfaction.

This brief discussion certainly doesn't settle the debate between compatibilists and incompatibilists. It would be gross hubris to think such a debate can be settled so easily. While my dissertation is a defense of compatibilism, I have not focused much attention on meeting incompatibilists on their own terms. The reason for this is that my aim has been to present what I take to be the best compatibilist theory one can give. I count it as a defense of compatibilism, and thus a defender of incompatibilism must find fault with my argument, if he is to defend incompatibilism. Incompatibilists must show either that my theory cannot explain important phenomena any theory of responsibility ought to explain, or that I'm wrong that determinism wouldn't invalidate one of my conditions. I cannot here assume that such an effort by an incompatibilist would fail, though I think one advantage of my strategy is that it places the onus on the incompatibilist again to tell us what my view has missed. Nevertheless, resolving the debate in one fell swoop is hindered in large part due to the resilience of the conflicting judgments held by compatibilists, on the one hand, and incompatibilists, on the other. Indeed, I think that the debate suffers from "dialectical stalemate[s]."[233] I have tried to address some of these in this dissertation.[234] As I have done throughout, however, I will chiefly ignore the long-standing traditional debate in this chapter, and target again the prospects of compatibilism itself. To conclude, then, I will instead focus on some possible future threats to responsibility, and why compatibilists should see such threats in an optimistic light.

To put it bluntly, I believe that the toughest threats to a defensible theory of responsibility will challenge those capacities of agency it seems we obviously possess. Any thesis that, if true, entails that we don't have one or any of the three corresponding capacities I've discussed would prove to be a serious challenge to the very possibility of responsibility. But the upshot of all this is actually good news for the compatibilist

[233] To use a phrase of John Martin Fischer's.

[234] For example, in my discussion of Ultimate Responsibility, Chapter 4, Section IV.4.

because it makes the cost of living without responsibility that much more dire, and, I think, easier to resist.

First, if the impossibility of responsibility is only achieved by a thesis suggesting that we lack one of the capacities listed above, then this would be a radical thesis indeed. We have already noted how humans lacking one of the above capacities would be aberrant; so a thesis that made such an aberration the norm would be striking and quite radical. Such a thesis would seem to be in need of stringent defense, and would face quite daunting argumentative hurdles.

Second, many hard determinists make the following general dialectical move.[235] While their view insists that we aren't responsible for what we do, they seek to show that the loss of responsibility is no great loss at all. We can retain much of what's important to us in life even if we are not truly responsible for what we do. Thus, they seek to minimize the costs associated with giving up responsibility. Such an avenue would seem to be blocked if the only route to hard incompatibilism there is is by showing we don't possess what I take to be basic agential capacities.[236] It would be a quite dire cost of giving up responsibility if it entailed that we can't explain action by reference to belief-desire sets or that we can't foresee the differences we make to the world. These are costs that are worth bearing in mind when considering how to weigh potential threats to the possibility of responsibility.

Third, the more radical a thesis, and the more dire its consequences, the easier it is to reject. Dialectically, the burden of argument rests on the stronger thesis, on the one that diverges most from our common understanding of things. This isn't to say such

[235] For specific examples, see Pereboom [2001], Sommers [2007].

[236] Of course, hard incompatiblists who make this move presumably think responsibility requires more than satisfaction of my three conditions, but that is an argument that can't be settled here.

burdens can't be met; of course they can. Quantum physics, for example, often flies in the face of our common sense understanding of the world, yet it has extraordinary explanatory and predictive power. But any thesis that challenges the possibility of responsibility must meet similar burdens, I think. For they will have to show that losing responsibility and some basic capacities of agents is worth what we gain from accepting the thesis. And while it is possible positions will come along that can meet this burden, I think compatibilists should have some optimism that such theses are not readily forthcoming.

So the serious threats to responsibility will come from radical theories like epiphenomenalism, which claims that mental states are causally inefficacious. Epiphenomenalism, it seems, would show that my conditions on responsibility can't be met.[237] Similarly, eliminativists like the Churchlands, who deny the reality of mental states, will also think that we lack the capacity to act on beliefs and desires, and therefore we can't meet my Voluntariness Condition.

Two points are in order here. First, requiring beliefs and desires to be causally implicated in action is my gloss on the Voluntariness Condition. The condition itself is meant to distinguish between those actions that we perform voluntarily, like opening a bottle of champagne or taking a coat, from those bodily movements that aren't voluntary, like sneezes and spasms. So if epiphenomenalism or eliminativism were true, we should still be able to distinguish between sneezes and bottle-openings, but we might need a new way to capture that distinction.

[237] In particular, Eddy Nahmias has been developing a recent charge that theses like epiphenomenalism pose a serious threat to responsibility.

Second, and more importantly, even if epiphenomenalism and eliminativism are threats to responsibility on my account, this is no *special* worry for my theory. These theses are threats to every account of responsibility I can think of. In the case of eliminativism, the implication is that we don't really have minds, and if this claim were true, it should come as no surprise that we're not responsible for what we do. In the case of epiphenomenalism, while we would have minds, they would be totally disconnected from our actions. If true, this thesis would stand as an objection to *all* extant theories of responsibility, and would imply that our minds play no role in issuing action. Such a thesis would pose difficulties for more than just accounts of responsibility. Nevertheless, both epiphenomenalism and eliminativism about the mental seem to me to be highly contentious theories, and thus of less concern for those who wish to secure responsibility than determinism has ever been thought to be. They are worries I think any compatibilist can live with, mostly because they seem highly implausible.

6.4. Are Animals Responsible?

THERE IS ONE FINAL worry I must address. One might object that if it is so easy to meet my conditions on responsibility, then perhaps certain non-human animals could satisfy them. But we don't think animals are responsible, so my account must be wrong if it has this implication. The worry is, essentially, that my account gives responsibility away too easily. So some condition or other must be missing.

It is true that we don't think clams and snails can be morally responsible. Nor do we blame viruses for illness, in any morally relevant sense. And we can work up from this level of biological sophistication: we don't think deer blameworthy for eating our shrubs, or bears for killing spawning salmon, and we wouldn't likely praise dolphins who rescue drowning swimmers. While some might blame or praise their dogs, it is more likely that this is seen in an educative light, as training the dog to behave certain ways,[238] then as distinctively moral practices. And, presumably, many would deny moral responsibility right up through higher-order primates, as well. The worry expressed above seems to suggest a line dividing human agents from other animals. It is a line that seems at home in the free will debate, given the large tradition of seeing free will as a distinctively human ability. This feature has been translated into moral responsibility talk, such that it seems as if only humans could be morally responsible for what they do.

At the same time, humans are surely animals. We are unique in certain respects: our capacities for language and problem-solving place us at the top of cognitively sophisticated animals. We're the only animals, for instance, who have debates about the prospects of responsibility (as far as we know). Yet there is also massive evidence that we aren't as special as we might think we are. Data has illustrated just how cognitively sophisticated chimpanzees and other great apes are; we know that dolphins are quite smart; and lots of animals display various sorts of behavior that seem to indicate various levels of complicated reasoning. Even rats seem to be able to reason causally about the effects of their actions.[239]

So, we might ask, how special are we? Are humans unique enough that only we could be responsible for what we do? Or is this just another instance of a false, but strongly held, conviction about responsibility? If a theory of responsibility implies that

[238] Similar to the way we might blame a non-responsible young child as a form of moral education.
[239] Dickinson and Shanks [1995].

animals can be responsible for what they do, is this a strike against it? Does such an implication rule out the theory?

Let me begin to address these questions by making a conjecture: some animals[240] possess the relevant agential capacities to at least as high a degree as children.[241] Coupling this conjecture with the assumption that children are responsible for at least some of what they do, it follows that some animals can be responsible as well. It is widely accepted that responsibility will be grounded, at least in part, on the capacities of agents. The first part of this chapter was devoted to my account of responsible agency: those capacities an individual must possess to ground the presumption that he is responsible for the particular things he does. If responsibility does depend on agency, then if an individual is determined to have the relevant agential capacities, he may well be responsible for particular things. Whether he is or not will depend on satisfying the conditions on responsibility, conditions that relate the individual to a particular object of responsibility. But if we deem A an agent (with the right sorts of capacities), then A is "in the ballpark" for ascriptions of responsibility.[242] If some animals are also at least in the ballpark regarding agency, then it seems to me no great worry that my account implies that they could be responsible. For it would be the case that since many animals have certain distinctive agential capacities, and since responsibility is grounded largely on agency, it should be no surprise that at least some animals turn out to be potentially responsible for things they do.

[240] Despite my initial comparisons, from here on out I'll use the term 'animal' in contradistinction to 'human'.

[241] For a discussion of this claim, see Gomez [2004].

[242] The use of "in the ballpark" I borrow from Fischer and Ravizza, who look to me to be giving primarily an account of responsible agency, not a theory of 'local' responsibility (which addresses responsibility for particular objects).

Nevertheless, I’m not convinced that animals would actually satisfy my conditions on responsibility, even if they have some basic agential capacities. Resolution of the issue depends again on empirical science and what it tells us about the abilities of animals. There does seem to be evidence that many animals act on belief-desire sets.[243] And we can certainly distinguish between when an animal apparently does something on purpose and when an animal does something as the result of a spasm or unconscious movement. So I’m willing to concede animals do things voluntarily, as I understand the condition. We could certainly draw the same distinction the Voluntariness Condition captures in human action in animal conduct as well.

It seems plausible that many animals can foresee some consequences of their actions. Means-end reasoning is sufficient for that. But notice that already the range of viable objects of responsibility begins to narrow here. One can only be responsible, on my view, for an outcome one foresaw as a possible consequence at the time of action. Presumably, the capacity to foresee in the manner required is less developed in most animals than in humans. Thus, the range of things animals could be responsible for diminishes as a result. However, I will concede that when a chimp uses a stick to get termites out of their mound, or a raccoon gets a trashcan lid off, they act intentionally. More to the point, they seem to satisfy the Intentionality Condition, and could be responsible in each case.

Let’s even suppose that on a plausible understanding of animal psychology, they can possess relevant factual concepts of what they are doing. I suspect that only higher-order animals will be able to thereby meet the No-Mistake Condition. For instance, it is possible that a mole believes that he is eating the roots of a delicious plant, but he

[243] [[[citations]]]

probably doesn't know that it is your plant. And having that belief is crucial to responsibility, as Jan's case with the coat-taking illustrated. But many animals might indeed possess concepts like 'pain' or 'damage,' and thus believe that a consequence of their action would be causing pain or damage. If so, it is possible that these facts would be sufficient for the relevant moral verdicts,[244] and thus animals could satisfy the No-Mistake Condition after all.

Indeed, we might still wonder about very sophisticated animals, like chimps and dolphins. Is a chimp blameworthy for purposefully killing another chimp?[245] Are dolphins who rescue shipwreck victims praiseworthy for their deeds?[246] I puzzle over these questions still. One stumbling block is that between humans we have the benefit of language. Our shared practices of blaming and praising are shared largely due to how language holds them together. Indeed, Michael McKenna has claimed that the framework of responsibility depends on communication.[247] I agree to a point. To hold each other responsible requires communicating our demands and expectations, of having these known by others. Holding others responsible requires being able to communicate our attitudes when they violate these expectations or surpass them. But, as I've argued throughout the dissertation, the important features of holding each other responsible are of derivative importance to the features of *being* responsible. So it could always be the

[244] Then again, perhaps our moral verdicts only range over human conduct and enterprises. I have no argument for why this would be the case, but if it were, then either animals would fail the No-Mistake Condition, or they could be responsible but never blameworthy or praiseworthy.

[245] There is evidence chimp groups do kill their own, as reported by primatologist David Watts. His hypothesis is that these "group kills" reinforce social bonds among the participating members. If true, this hypothesis might match sociological explanations of gang behavior among humans. Indeed, we might expect this to be the case given the similarities between humans and chimpanzees.

[246] I don't know the prevalence or accuracy of such reports. But we can examine the question as a hypothetical nonetheless.

[247] See McKenna [1998].

case, on my view, that some animals are responsible, even though there's no sense in holding them responsible.

Here, again, we have the distinction between an individual being responsible for something and it's being the case that we ought to blame him. It might be the case that at least some animals (maybe many) are responsible for (some of) what they do, it just doesn't make any sense to hold them responsible, since it would have no effect on behavior. This thought really isn't as strange as it might initially sound. The notion of responsibility, as I noted in the Introduction to this dissertation, has been developed as a central aspect to a vast array of *human* relationships and social practices. Responsibility, the fact of an individual's being responsible, has a distinctly human importance. It lacks this importance in our relationships with animals, and in turn animals lack the practices and relationships between each other that would require such a notion. So it would certainly seem that even if animals were actually responsible for some of what they do, this fact would not imply that we ought to go about blaming or praising them.

Nevertheless, some of us humans do blame and praise animals. Those sportswriters who suggested Secretariat's performance in the Triple Crown merited inclusion in "The 50 Top Athletes of the 20th Century"[248] were, deliberately or not, praising Secretariat for his performance. And much of the talk about the way he raced, the interpretation of his actions, suggest a commitment to his acting very much like a human athlete, possessing the same admirable qualities (at least with respect to athletic performance). Similarly, a friend of mine noted in discussion that he and his wife do in fact blame their dog for transgressions of behavior. He said that the dog knows she isn't allowed to do certain things, and when she does them she deserves certain forms of

[248] ESPN [1999].

sanctioning treatment. In this case, it seems to me, my friend is assuming that he and his dog do share a network of commitments and expectations, and are participating in a shared practice that does depend on the importance of responsibility. Of course, his understanding of the practice may be much different than his dog's, but he saw no problem in holding his dog responsible for her conduct (at least in specific cases).

This anecdotal evidence is not meant to prove the point, only to show that reluctance to accept the implication that animals can be responsible can be mitigated by reflection on our experiences with those animals that hold special significance to us.

So, why not think the chimp blameworthy for killing, and the dolphin praiseworthy for saving? The worry may be that these animals are not really deliberating when they act. They just behave according to nature's programming; they don't choose. But this is a dangerous path to start down. For we mustn't forget a simple truth: all humans are animals. And we are programmed by nature as much as chimps and dolphins, or so it seems. The general conclusion, then, is to the degree that other animals come close to the cognitive sophistication of humans (at least regarding the three capacities discussed in this chapter), I think they approach the capacity to be responsible for things they do.

This is not a popular position; but then again, the question of animal responsibility is not much discussed. It seems to me that most assume only human agents can be responsible, and this is because responsibility is so tightly connected to morality and only humans are in the business of morality. But it's hard for me not to see the chimps' action as expressing an ill quality of will towards the other chimp, and the dolphins expressing good will towards the shipwreck crew. I don't take this to be evidence in favor of either actually being responsible. I merely mention these appearances to highlight my initial response to the worry: given that so much of the normative domain, especially things like responsibility and morality, depends on our nature as agents, then it wouldn't be surprising to find that those animals who are most similar to us as agents should be "in the ballpark" for being responsible. Perhaps dolphins are praiseworthy for rescues. Perhaps they are not. In any case, I submit that the answer to the question depends on whether they are agents like we are. And unless one thinks that because dolphins don't go into the business of praising they can't be praiseworthy, I don't think an affirmative answer would be all that striking or worrisome. It certainly wouldn't imply that we should actually start praising dolphins. Instead it would suggest that perhaps humans do not have the exclusive rights (pardon the pun) to the business of morality.[249]

6.5. Conclusion

COMPATIBILISM OFFERS US the most promising theory of responsibility, and I have argued for what I take to be the most promising compatibilist account. It secures for us responsibility for much (indeed, most) of what we do, and so even should science tell us the world is thoroughly deterministic, we can still retain the robust notions of friendship and love, the complicated social practices of blaming and praising, and our genuine sense of accomplishment at artistic creation or scholarly achievement, all of which depend on our being responsible for what we do.

[249] To be clear, all that I would require for dolphins to be in the business of morality is that they see something good in saving the shipwrecked crew. More would have to be said to fully work out the details here, but I don't find it too implausible a suggestion. And, in any case, I certainly wouldn't consider such a result worrisome for a theory of responsibility.

My particular brand of compatibilism also has the advantage of securing responsibility given the satisfaction of some fairly uncontroversial conditions and capacities, making it all the more likely that we satisfy them, and making my compatibilism that much more defensible in light of potential future threats. Indeed, as easy as it is to meet my conditions on responsibility, it seems the likely threats to my account will come from less and less plausible positions in action theory and philosophy of mind.

I've also shown that issues that currently worry compatibilists needn't worry us. Normative competence, manipulation, and the ability to do otherwise, don't pressure us away from a thoroughly compatibilist program. The compatibilist has resources to explain just why these issues *seem* problematic, but those same resources help explain why they needn't remain impediments to a simple and explanatorily powerful compatibilist position.

My goal from the start has been to defend an account of responsibility that explained what was essentially important to our lives as human beings. What is crucial about being responsible for what we do, and what must be true of us and our actions to secure that notion? My answer is simple. To be responsible for an action or outcome, it must be the case that: (1) the outcome is the result of an action explainable by a belief-desire set; (2) the outcome was at least a foreseen possible effect of that action; and, (3) the agent had no false beliefs about the nature of his action or outcome necessary for generating an evaluation by the relevant normative standards. Satisfying these conditions is necessary and sufficient for being responsible for a particular outcome. If the outcome is bad, the agent is blameworthy. If the outcome is good, the agent is praiseworthy. Responsibility is primarily a relation tying individuals to the things they do or bring about, such that the individual is evaluable by that action or outcome's lights. To be so evaluable, my three conditions must be met, and they are quite easily met even if the world should be thoroughly deterministic. Indeed, I believe my three conditions can be met even should the world be partially indeterministic. In any case, I think nothing more than my three conditions is required to be responsible, though one is free to challenge my claim.

References

Austin, J. L. (1957). "A Plea for Excuses." Proceedings of the Aristotelian Society **57**: 1-30.

Bennett, J. (1980). Accountability. Philosophical Subjects: Essays Presented to P.F. Strawson. Z. v. Straaten, Oxford University Press.

Berofsky, B. (2000). "Ultimate Responsibility in a Deterministic World." Philosophy and Phenomenological Research **60**(1): 135-140.

Brand, M. (1984). Intending and Acting: Toward a Naturalized Action Theory, The MIT PRess.

Clarke, R. (2004). Incompatiblist (Nondeterministic) Theories of Free Will. Stanford Encyclopedia of Philosophy.

Clarke, R. (2005). "On an Argument for the Impossibility of Moral Responsibility." Midwest Studies in Philosophy **29**: 13-24.

D'Arms, J. and D. Jacobsen (2000). "The Moralistic Fallacy: On the 'Appropriatenes' of Emotions." Philosophy and Phenomenological Research **61**(1): 65-90.

D'Arms, J. and D. Jacobsen (2000). "Sentiment and Value." Ethics **110**: 722-748.

Dennett, D. (1984). Elbow Room: The Varieties of Free Will Worth Wanting, The MIT Press.

Dickinson and Shanks (1995). Instrumental Action and Causal Representation. Causal Cognition. D. Sperber, D. Premack and A. Premack, Oxford University Press.

Ekstrom, L. W. (2000). Free Will: A Philosophical Study, Westview Press Focus Series.

ESPN (1999). The Top 50 Athletes of the 20th Century.

Finkelstein, C. (2005). "Responsibility for Unintended Consequences." Ohio State Journal of Criminal Law **2**(579): 579-599.

Fischer, J. M., R. Kane, et al. (2007). Four Views on Free Will, Wiley-Blackwell.

Fischer, J. M. and M. Ravizza (1992). "Responsibility, Freedom, and Reason." Ethics **102**: 368-389.

Fischer, J. M. and M. Ravizza (1998). Responsibility and Control, Cambridge University Press.

Fischer, J. M. and N. Tognazzini (forthcoming). "The Truth About Tracing." Noûs.

Frankfurt, H. (1971). "Free Will and the Concept of a Person." Journal of Philosophy **68**: 5-20.

Frankfurt, H. (1988). The Importance of What We Care About, Cambridge University Press.

Frankfurt, H. ([1969], 2003). Alternate Possibilities and Moral Responsibility. Moral Responsibility and Alternative Possibilities. D. Widerker and M. McKenna, Ashgate**:** 17-21.

Gomez, J. C. (2004). Apes, Monkeys, Children, and the Growth of the Mind, Harvard University Press.

Graham, P. (ms 1). "A Theory of Blameworthiness."

Graham, P. (2005). Blame, Determinism, and Ignorance. Department of Philosophy. New York City, New York University. **Ph. D.**

Greenspan, P. (1988). Emotions and Reasons: An Enquiry Into Emotional Justification, Routledge, Chapman, and Hall.

Greenspan, P. (2003). "Responsible Psychopaths." Philosophical Psychology **16**(3): 417-429.

Haji, I. (2003). "Determinism and Its Threats to the Moral Sentiments." The Monist **86**(2): 242-260.

Hart, H. L. A. (1968). Punishment and Responsibility: Essays in the Philosophy of Law, Oxford University Press.

Kane, R. (1996). The Significance of Free Will, Oxford University Press.

Kane, R., Ed. (2001). The Oxford Handbook of Free Will, Oxford University Press.

Kane, R. (2005). A Contemporary Introduction to Free Will, Oxford University Press.

Kant, I. ([1788], 2002). The Critique of Practical Reason, Hackett.

Knobe, J. (2003). "Intentional Action and Side Effects in Ordinary Language." Analysis **63**: 190-193.

Knobe, J. (2004). "Intention, Intentional Action and Moral Considerations." Analysis **64**: 181-187.

McKenna, M. (1998). "The Limits of Evil and the Role of Moral Address: A Defense of Strawsonian Compatibilism." The Journal of Ethics **2**: 123-142.

McKenna, M. (2007). "Putting the Lie on the Control Condition for Moral Responsibility." Philosophical Studies.

Mele, A. (2006). Free Will and Luck, Oxford University Press.

Montmarquet, J. (2002). "Wallace's 'Kantian' Strawsonianism." Philosophy and Phenomenological Research **64**(3): 687-692.

Naylor, M. B. (1985). "Voluntary Belief." Philosophy and Phenomenological Research **45**: 427-436.

"Notes" (1972). "Negligence and the General Problem of Criminal Responsibility." The Yale Law Journal **81**: 949-979.

Oshana, M. (1997). "Ascriptions of Responsibility." American Philosophical Quarterly **34**(1): 71-83.

Pereboom, D. (2001). Living Without Free Will, Cambridge University Press.

Petit, G. (2002). "Are We Rarely Free? A Response to Restrictivism." Philosophical Studies **107**: 219-237.

Rabinowicz, W. and T. Rønnow-Rasmussen (2004). "The Strike of the Demon: On Fitting Pro-Attitudes and Value." Ethics **114**: 391-423.

Rodin, D. (2002). War and Self-Defense. Oxford, Clarendon University Press.

Rosen, G. (2002). "Culpability and Ignorance." Proceedings of the Aristotelian Society **103**: 61-84.

Sartre, J.-P. ([1943], 2003). Being and Nothingness. The Philosophy of Jean-Paul Sartre, Vintage Books.

Shemmer, Y. (2004). "Desiring at Will and Humeanism in Practical Reason." Philosophical Studies **119**(3): 265-294.

Smart, J. J. C. (1961). "Free-Will, Praise, and Blame." Mind **70**(279): 291-306.

Smilansky, S. (2007). "Determinism and Prepunishment: The Radical Nature of Compatibilism." Analysis **67**(4): 347-349.

Smith, A. (2005). "Responsibility for Attidues: Activity and Passivity in Mental Life." Ethics **115**: 236-271.

Sommers, T. (2007). "The Objective Attitude." The Philosophical Quarterly **57**(228): 321-342.

Strawson, G. (1994). "The Impossibility of Moral Responsibility." Philosophical Studies **75**(1-2): 5-24.

Strawson, G. (2001). The Bounds of Freedom. The Oxford Handbook of Free Will. R. Kane, Oxford University Press**:** 441-461.

Strawson, P. F. (1962). Freedom and Resentment. Proceedings of the British Academy. **68:** 1-25. Reprinted in Watson, G. (ed.). Free Will. Oxford University Press, 1982, pp.59-80.

van Inwagen, P. (1975). The Incompatibility of Free Will and Determinism. Philosophical Studies. **25:** 185-199.

van Inwagen, P. (1989). When is the Will Free? Philosophical Perspectives, 3, Philosophy of Mind and Action Theory. J. E. Tomberlin, Ridgeview.

van Inwagen, P. (1994). "When the Will Is Not Free." Philosophical Studies **75**: 95-113.

Vander Laan, D. (2001). "A Regress Argument for Restrictive Incompatibilism." Philosophical Studies **103**: 201-215.

Vargas, M. (ms 1). "Building a Better Beast."

Vargas, M. (2004). "Responsibility and the Aims of Theory: Strawson and Revisionism." Pacific Philosophical Quarterly **85**(2): 218-241.

Vargas, M. (2005). "The Trouble With Tracing." Midwest Studies in Philosophy **XXIX**: 269-291.

Vihvelin, K. (2000). "Freedom, Foreknowledge, and the Principle of Alternate Possibilities." The Canadian Journal of Philosophy **30**(1): 1-23.

Wallace, R. J. (1994). Responsibility and the Moral Sentiments, Harvard University Press.

Watson, G. (1982). Free Will, Oxford University Press.

Watson, G. (1987). Responsibility and the Limits of Evil: Variations on a Strawsonian Theme. Responsibility, Character, and the Emotions. F. Schoeman, Cambridge University Press.

Widerker, D. and M. McKenna, Eds. (2003). Moral Responsibility and Alternative

Possibilities: Essays on the Importance of Alternative Possibilities, Ashgate Publishing.

Williams, B. (1973). Deciding to Believe. Problems of the Self, Cambridge University Press.

Winters, B. (1979). "Believing at Will." Journal of Philosophy **76**: 243-256.

Wolf, S. (1980). "Asymmetrical Freedom." The Journal of Philosophy **77**(3): 151-166.

Wolf, S. (1987). Sanity and the Metaphysics of Responsibility. Responsibility, Character, and the Emotions. F. Schoeman, Cambridge University Press.

Zimmerman, M. (1986). "Negligence and Moral Responsibility." Noûs **20**: 199-218.