RESPONSIBILITY AND MANIPULATION

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Charles Zachary Cogley, M.A.

Philosophy Graduate Program

The Ohio State University

2010

Dissertation Committee:

Daniel Farrell, Advisor

Justin D’Arms

Don Hubin
ABSTRACT

In my dissertation, I defend compatibilism against the problem presented by manipulated agents. I argue that compatibilists about moral responsibility and determinism should claim that people whose attitudes are manipulated by outside interveners are nonetheless fully morally responsible for what they do, so long as such people satisfy compatibilist conditions on being morally responsible.

To show this, I demonstrate that our practice of holding each other morally responsible has two distinct components, one emotional, the other non-emotional. The emotional component consists of a class of reactive emotions that are made appropriate by the actions and attitudes of other people. The non-emotional component concerns the way our trust of others is properly impacted by their actions and attitudes.

I develop a full account of both the emotional and non-emotional components of holding each other morally responsible, and I also go on to characterize the circumstances under which our emotional and non-emotional attitudes towards others and their actions are appropriately modified. I then apply this framework to cases of manipulated people and show that when the people in question satisfy compatibilist accounts of free agency, their manipulation should not impact our willingness to hold them responsible. Finally, I show that, within certain limits, people who commit crimes because they have been manipulated in this way may be permissibly punished.
The ultimate payoff is threefold: a more nuanced understanding of the situations in which people are legitimately held morally responsible for what they do, the development of a compatibilist account of moral responsibility that is immune to many of the canonical objections, and a novel theory of punishment.
DEDICATION

Dedicated to Andrea, Phil and my mother and father
ACKNOWLEDGEMENTS

I thank my advisor, Dan Farrell. Dan is my model for philosophical interaction. He put up with countless revisions with good cheer and this project is immeasurably better for his input. I thank Justin D’Arms. Justin provided the conceptual spark that began this project and helped me hash out the argument of one of the key chapters. I dreaded metaethics until I learned it from him. I thank Don Hubin. I’ve learned an amazing amount for Don’s classes and while I doubt that I’ll ever be as organized or of generally good cheer, he provides an exemplar to aspire to.

I owe a special debt to Pamela Hieronymi. I learned an immense amount from Pamela and she always treated me as a full colleague. She is incredibly generous, both personally and professionally, and I’m deeply grateful.

I am grateful for the friendship and encouragement of many of my fellow graduate students, particularly Erin Brown, Steve Brown, Eric Carter, Andy Choi, Julian Cole, Salvatore Florio, Dai Heide, Nick Jones, Alison Kerr, Matt Jordan, David Merli, James Macpherson, Bill Melanson, Ryan Nichols, Henry Pratt, Cate Sherron, Josh Smith, Bill Roche, Kelly Trogdon and Cathal Woods. I am particularly indebted to Jesse Duarte, Timmy Fuller, Carol Hay and Dai Heide for their intellectual and moral support.

Finally, I thank Andrea Scarpino for her incredible support. She read every page of this work and believed it would be finished even when I didn’t. Without her involvement I doubt I would have completed this project.
VITA

1998...........................................BA, Philosophy, Saint Louis University

2001...........................................MA, Philosophy, University of Cincinnati

2001-2006.................................Graduate Teaching Associate,
                               The Ohio State University

2006-2008.................................Lecturer, University of California, Los Angeles

FIELDS OF STUDY

Major Field: Philosophy
# Table of Contents

Abstract....................................................................................................................................ii  
Dedication....................................................................................................................................iv  
Acknowledgements.......................................................................................................................v  
Vita................................................................................................................................................vi  
Chapter 1: Introduction..................................................................................................................1  
Chapter 2: Reactive Emotions and Moral Responsibility..............................................................20  
Chapter 3: Trust............................................................................................................................53  
Chapter 4: Reactive Nonemotions and Moral Responsibility.......................................................83  
Chapter 5: Distrust, Desert and Punishment...............................................................................110  
Bibliography................................................................................................................................144
1.1 Wells and Plum

Consider the case of Brian Wells. On August 28, 2003, Brian Wells used a shotgun to rob a bank in Erie, Pennsylvania. Shortly afterward, Wells was apprehended by police, who discovered that a large metal collar with a bomb attached to it was locked around his neck. The police also discovered that a group of people (call them “the Criminals”) had given Wells a set of instructions that required him to rob the bank. If Wells did not carefully follow the Criminals’ instructions, the bomb locked around his neck would be detonated. The bomb squad was called, but arrived too late. The Criminals exploded the bomb, punching a softball-sized hole in Wells’ chest.¹

Would it be reasonable to say that Wells was morally responsible for robbing the bank? Though we would normally think that someone who robbed a bank was morally responsible for doing so, most people would question Wells’ moral responsibility because he was coerced into robbing the bank out of fear for his life. We would also question the appropriateness of punishing him (if he were still alive). Instead, the Criminals who

¹ For more details on the case, see John Caniglia, “Erie bombing 'victim' was in on bank robbery,” n.d., sec. Metro, http://blog.cleveland.com/metro/2007/07/man_forced_to_rob_bank_with_a.html. Accessed 4/20/2008. Wells apparently believed that the bomb would be fake and was in on the initial plot. However, when his collaborators informed him the bomb was real, he tried to escape, only to be apprehended and have the device locked to his neck. The fact that Wells was in on the plot initially may make some readers think he is morally responsible for robbing the bank. Therefore, fix your attention on a version of the case where Wells was not ever in on the plot.
locked the explosives to Wells’ neck were responsible for robbing the bank, and they should be punished if it is possible to apprehend them.

Now consider Derk Pereboom’s imaginary case concerning Professor Plum, who kills Ms. White for the sake of some personal advantage.

Professor Plum was created by neuroscientists, who can manipulate him directly through the use of radio-like technology, but he is as much like an ordinary human being as is possible given this history. These neuroscientists manipulate him to undertake the process of reasoning by which his desires are brought about and modified. They do this by pushing a series of buttons just before he begins to reason about his situation, thereby causing his reasoning process to be rationally egoistic. Plum does not think and act contrary to character since his reasoning process is often manipulated to be rationally egoistic. His effective first-order desire to kill White conforms to his second-order desires. The process of deliberation from which his action results is reason-sensitive; in particular, this type of process would have resulted in his refraining from killing White in some situations in which the egoistic reasons were otherwise. Still, he is not exclusively rationally egoistic, since he typically regulates his behavior by moral reasons when the egoistic reasons are relatively weak – weaker than they are in the current situation. He is also not constrained in the sense that he does not act because of an irresistible desire – the neuroscientists do not provide him with a desire of this kind.²

Just as in Wells’ case, most people will probably say that Plum is not morally responsible and should not be punished. While Plum was not controlled by the neuroscientists in the same way that Wells was coerced by the criminals, the fact that Plum was manipulated by the neuroscientists will lead most people—and most philosophers—to deny that Plum is morally responsible for robbing the bank.

1.2 Constraining Control

Cases like Plum’s present a problem for philosophers like myself who are compatibilists about moral responsibility and determinism. We believe that even if all human actions turn out to be casually determined, human beings will still be morally responsible for at least some of the things we do. Other philosophers, called incompatibilists, deny this. They believe that if all human actions are causally determined, then human beings are never morally responsible for what they do.

The debate between compatibilists and incompatibilists concerns the nature of the kind of control necessary for a person to be morally responsible for what she does. Both sides accept that in order for a person to be morally responsible for something, that person must exhibit a kind of control over her actions. Compatibilists claim that causal determinism is no threat to human beings satisfying the control condition on moral responsibility. Incompatibilists believe that causal determinism rules out people being morally responsible because determinism makes the necessary kind of control impossible.

To begin to understand why Plum’s case presents a problem for compatibilists, let us first return to Wells for a moment. Wells’ case presents an example of what Robert

---

3 I am here using “compatibilism” and “incompatibilism” to refer to the compatibility question with respect to moral responsibility and determinism. These terms can also be used to indicate positions on the compatibility of free will and determinism, where “free will” is the power to act other than as one acts. See van Inwagen (1983), 8.

Kane has called constraining control (CC). When a person undergoes constraining control, she is forced to do something against her will. The controller in a CC case (in this case, the Criminals) gets what she wants by creating impediments to a person’s desired course of action. As I said above, most people would not take Wells to be responsible for robbing the bank because he was coerced by forces outside his control, or perhaps because in these circumstances—when one’s life is threatened—it is not wrong to cooperate with The Criminals. A compatibilist who wants to explain and make sense of this responsibility attribution will want her theory to have it that people like Wells, who are manipulated by CC controllers, are not responsible for the actions that resulted from this control.

1.3 Covert Nonconstraining Control

While it may be easy for a compatibilist to show that her theory will deny that Wells is morally responsible, it will be much harder for a compatibilist to get the intuitively correct answer in Plum’s case. Plum’s case presents an example of what Kane calls covert nonconstraining control (CNC). As Kane notes, CNC is epitomized by cases of brainwashing and other kinds of manipulation where the controllers get their way by manipulating a person’s agential structure directly, and secretly. This presents a

---

5 The Significance of Free Will (New York: Oxford University Press, USA, 1996), 64-65.
6 I prefer the latter explanation because, strictly speaking, Wells was not “compelled” to rob the bank. However, at this point I am operating at an intuitive level in order to fix the problem.
7 By a person’s “agential structure” I mean whatever mental attitudes and operations are the cause of that person’s action. A person’s agential structure could include beliefs, desires, reasons, emotions, intentions, etc. A simple model of action might start with a person’s desire, D, that she knows will lead to some result, R, that she wants to bring
problem for compatibilists who hold that only certain agential features must be satisfied in order for someone to be responsible.

By a person’s “agential structure” I mean whatever mental attitudes and operations are the cause of that person’s action. A person’s agential structure could include beliefs, desires, reasons, emotions, intentions, etc. A simple model of action might start with a person’s desire that she knows will lead to some result, R, that she wants to bring about. She also believes that doing X, some action, will lead to R. If the person’s belief and desire together cause her to perform X in order to bring about R, then the belief and desire constitute what I mean by that person’s “agential structure.”

In fact, Pereboom constructed Plum’s case so it satisfies four of the most influential compatibilist accounts of the relevant kind of agential structure on offer, those proposed by David Hume, Harry Frankfurt, John Martin Fischer and Mark Ravizza, and R.J. Wallace. The Humean condition is that Plum does not think and act contrary to character; the Frankfurtian condition concerns the mesh between Plum’s first- and second-order desires; the condition regarding the reasons-responsive process is due to Fischer and Ravizza; and the condition proposed by Wallace concerns Plum’s capacity to generally regulate his behavior by moral reasons. Plum thus satisfies an impressive litany of conditions that an influential batch of compatibilists have individually proposed as sufficient for moral responsibility.

---

about. She also believes that doing X, some action, will lead to R. If the person’s belief and desire together cause her to perform X in order to bring about R, then the belief and desire constitute what I mean by that person’s “agential structure.”

8 The Significance of Free Will, 65.
The proposed compatibilist conditions have been satisfied, and yet most people would still say that Plum is not morally responsible. The case of Plum, then, can be the basis of an argument that all these compatibilist views are incorrect. If these views are supposed to capture most people’s intuitions about when people have satisfied conditions sufficient for moral responsibility, then they fail miserably with Plum’s case. Perhaps, then, they should be rejected.

1.4 MANIPULATION ARGUMENTS

Examples like Plum’s case are the basis of manipulation arguments that seek to show that compatibilist accounts of the control condition on moral responsibility fail. Plum’s case can provide the basis of an argument against compatibilism in the following way: all of the above compatibilist conditions are satisfied in Plum’s case, and yet it intuitively seems as if he is not morally responsible for killing White. This suggests that these conditions are not sufficient for moral responsibility. But with a little thought we could construct a case like Plum’s for any compatibilist account of the control condition on moral responsibility.\(^9\) If such a case can be constructed for any compatibilist account, this suggests that compatibilism has a deep flaw that prevents it from getting the intuitively correct result in CNC cases like Plum’s. Therefore, it seems we should reject compatibilism.\(^{10}\)

---


\(^{10}\) Not all manipulation arguments have such a wide scope, of course. Most manipulation arguments target only one compatibilist account of the conditions sufficient for moral responsibility. But one of my claims is that manipulation cases pose potential problems
In spite of the fact that many people—including most philosophers—will find this counterintuitive, I believe Plum is morally responsible and in this dissertation, I defend that claim. This view about manipulation cases like Plum’s has been called hard compatibilism by Robert Kane.\(^\text{11}\) In taking the hard line, I join Michael McKenna, who has defended this view in three recent papers.\(^\text{12}\)

In the next section, I’ll discuss McKenna’s argument in more detail, but first, here is a brief methodological guide. Most philosophers working on moral responsibility hold that what we are doing when we hold someone morally responsible should be understood in a Strawsonian vein.\(^\text{13}\) Few papers loom as large over a particular literature as

---

for any compatibilist account. Derk Pereboom’s “Four Cases” argument is one manipulation in the literature that does aim to trouble any possible compatibilist account. Much more on Four Cases in a later section.

Saul Smilansky lists manipulation arguments as one of three fundamental issues that push him toward skepticism about free will and moral responsibility, indicating the fact that many current theorists see manipulation as being a problem for all compatibilist accounts of moral responsibility and free will. He writes, “The second challenge to compatibilism is more internal, and begins with our intuitions on specific cases (such as instances of gross manipulation). Those cases, which we all see to be tainted, are then shown, through minute modifications, not to be fundamentally different from the cases that the compatibilist takes to be models of control and responsibility.” “Fischer’s Way: The Next Level,” *The Journal of Ethics* 12, no. 2 (June 3, 2008): 151.

\(^\text{11}\) *The Significance of Free Will*, 67.


\(^\text{13}\) I use the term “Strawsonian” to refer to philosophers who adopt a conception of responsibility that bears affinity with the one Strawson develops in “Freedom and Resentment.” Examples of prominent Strawsonians include Jonathan Bennett, “Accountability,” in *Philosophical Subjects: Essays Presented to P.F. Strawson*, ed. Zak...
Strawson’s “Freedom and Resentment” does over the literature on moral responsibility. While there is significant dispute over exactly what to make of Strawson’s argument there (and later I will broach some interpretive issues), there is no dispute that, since Strawson, most philosophers working on moral responsibility have supposed that the concept of moral responsibility can be understood by reference to the appropriateness of a set of attitudes and reactions Strawson calls the “reactive attitudes,” responses such as “gratitude, resentment, forgiveness, love, and hurt feelings.”

Given that it is generally agreed that we should understand a person’s moral responsibility in this way, here is a guide to how I will use several related notions. To think that someone is morally responsible is to think that having a particular attitude, or set of attitudes, toward that person is appropriate. For someone to be morally responsible is for a particular attitude, or set of attitudes, toward that person to be appropriate. To hold someone morally responsible is to have the attitudes toward that person that would be appropriate if that person were responsible. To hold someone accountable is to engage in behaviors or actions toward a person that are the natural or conventional expressions of holding that person morally responsible. A person is accountable when


some of these natural or conventional expressions of holding that person morally responsible are appropriate.

Also, I will assume in what follows that the associated notions of thinking, being and holding someone morally responsible (as well as accountable) concern more than just actions that are morally wrong or right, or morally good or bad. To see why I make this assumption, note the similarity in attitudes and reactions someone might have toward someone who cheated in the course of playing cards and someone who cheated on their taxes (assume that no moral norm was violated by the card-cheating. Both people did something “wrong,” but the person who cheated at cards violated a norm that is part of a game, while the tax-cheater violated a norm that is a legal (and perhaps moral) standard. Even though the standards the person violated were different, it would not be surprising if in both cases you would be angry with the violator, think that your anger is appropriate, intend to be careful when interacting with the person in the future, and then actually avoid the person.

In what follows, I will be counting these responses as instances of thinking and holding morally responsible (as well as accountable), regardless of whether the responses are to behavior that is evaluated by moral or nonmoral norms. The strength of our reactions to people tends to vary depending on whether the violation is one that we take seriously or whether the other person is someone important to us, but nothing ensures that these dimensions mirror distinctions between moral and nonmoral norms. A person’s moral responsibility for her actions do not depend on her actions being assessable by a moral norm.
1.6 McKenna’s Hard-Line Reply

In a trio of recent papers, Michael McKenna has sought to defend compatibilism from the threat posed by Pereboom’s case of Professor Plum. McKenna’s argument against Pereboom asks us to first consider a case of normal human action like this one:

Physicalist determinism is true, everything in the world is completely physical, and Plum is an ordinary human being, raised in normal circumstances, who is often but not exclusively rationally egoistic [just as egoistic as in the original case]. Plum’s act of killing White results from his undertaking the reasons-responsive process of deliberation, and he has the specified first- and second-order desires. He also possesses the general ability to grasp, apply, and regulate his behavior by moral reasons, but in these circumstances the egoistic reasons weigh very heavily for him, and as a result he is causally determined to murder White. But it is not due to an irresistible desire that he kills her.

McKenna then asks us to focus our attention, in this “normal” case, on the fact that in this version, Professor Plum is revealed “to satisfy a very rich picture of compatibilist free agency, one with a proper history, and in which Plum is ‘morally articulate’ in Strawsonian fashion.” Further, McKenna asks us to focus on the “important moral fact that Plum does kill Ms. White for egoistic reasons.” McKenna argues that these agential and moral features of the case support the claim that it is not evident in the normal case that Plum is not morally responsible. If it is true that in the normal case it is not evident that Professor Plum is not morally responsible, then

---

15 McKenna, “A Hard-line Reply to Pereboom's Four-Case Manipulation Argument.” McKenna’s argument in this paper builds on his work on manipulation cases from two earlier papers. McKenna, “Responsibility and Globally Manipulated Agents”; McKenna, “The Relationship Between Autonomous and Morally Responsible Agency.”

16 The text in the braces is mine. The case if from Pereboom, “Hard Incompatibilism,” 97.

17 “A Hard-line Reply to Pereboom's Four-Case Manipulation Argument,” 151. I’ll consider a Strawsonian view in the next chapter.

18 Ibid., 152.
McKenna will go on to argue by parity of reasoning that we should think the same of Plum in the manipulation case.

McKenna’s argument thus involves putting his own intuitions up against Pereboom’s and arguing, via Pereboom’s own methodology, that there is no relevant difference between the normal and manipulated versions of the case. Pereboom has the intuition that Plum in the manipulated case is not morally responsible and therefore also thinks that Plum in the normal case is not morally responsible. McKenna, on the other hand, has the intuition that Plum in the normal case may well be morally responsible and that, therefore, Plum in the manipulated case may well be morally responsible, too. If he can argue that the situation with Pereboom ends in a deadlock, McKenna will be satisfied with blunting Pereboom’s argument—after all, Pereboom began the argument. McKenna says, “If the compatibilist can merely cast doubt on the incompatibilist’s positive claim, then she has adequately replied to the argument even if she has not offered a positive defense of her thesis regarding manipulation.”

I want to register my qualified agreement with McKenna. It is true that if Pereboom’s intuitive argument has been adequately countered, then the compatibilist has done all she needs to do with respect to this one argument. There are many arguments that must be made in order to completely defend compatibilism and it would be unfair to require McKenna to offer all relevant arguments in order to rebut this one claim. Further, I think that McKenna is right to emphasize the agential and moral features of Plum in both the normal and manipulated cases in order to claim that Plum, in both cases, may

---

19 Ibid., 155. Italics original.
well be morally responsible. These are the right things to focus on to push back at Pereboom.

1.7 A FULL REPLY TO THE PLUM CASE

In spite of my appreciation for McKenna’s argument, I think compatibilists can reasonably aim higher than McKenna does. To indicate where I disagree with McKenna’s assessment of the dialectic, let me distinguish between McKenna’s *adequate* reply to Pereboom and what we might call a *full* reply. McKenna’s reply counts as adequate because it meets Pereboom’s intuitive argument with a contrasting intuitive argument of its own. If Pereboom’s argument rests entirely on the appeal to intuitions, then generating a forceful set of contrasting intuitions constitutes a respectable reply.

That is not all there is to Pereboom’s argument, however. He does not just supply a set of intuitions, but situates those intuitions within a Strawsonian account of moral responsibility that supports them. According to Pereboom,

> For an agent to be morally responsible for an action in the sense at issue in this debate is for it to belong to him in such a way that he would deserve blame if he understood that it was morally wrong, and he would deserve credit or perhaps praise if he understood that it was morally exemplary, supposing that this desert is basic in the sense that the agent would deserve the blame or credit just because he has performed the action, given understanding of its moral status, and not by virtue of consequentialist or contractual considerations.²⁰

This account of what it is for an agent to be morally responsible supports Pereboom’s intuitive argument by making sense of Pereboom’s intuitions about Plum. Since

---

Pereboom takes Plum in the manipulation case to intuitively not deserve blame, Plum is not morally responsible on Pereboom’s account of moral responsibility.

What would constitute a full reply to Pereboom would be for McKenna not just to offer his contrasting intuitions about the cases, but to situate those intuitions within a conception of moral responsibility that supports them. Or alternatively, McKenna might defend a conception of moral responsibility that instead is able to explain away Pereboom’s intuitions about the cases.

This is a point McKenna recognizes in his work on other agents subject to CNC manipulations. Surprisingly, however, it is not something he explores in the current dialectic with Pereboom. In his paper “Responsibility and Globally Manipulated Agents,” McKenna notes that rather than simply being on the defensive against cases of manipulation, a compatibilist might offer a positive argument in support of her own intuitions. In that paper, McKenna briefly characterizes an account of moral responsibility that, like Pereboom’s, ties moral responsibility closely to praise and blame. But rather than explaining moral responsibility in terms of praise and blame in Pereboom’s basic desert sense, McKenna gestures toward an account of moral responsibility built on what he calls a “conversational” model.21

21 I wish I could say more about what McKenna’s “conversational model comes to, but he only develops it through metaphor: “[H]olding another morally responsible for doing morally wrong is a manner of communicating with her. In particular, it is a manner of responding to what she has done as on an analogy with a conversation in which the blameworthy person’s conduct has a significance. This significance makes certain responses as opposed to others more fitting or meaningful. […] Her response to those holding her responsible, in terms of apology, a defiant unwillingness to acknowledge the wrong, a justification, a plea for forgiveness, and so on, is her further participation in the conversation.” “Responsibility and Globally Manipulated Agents,” 187-89.
I won’t spend time here interrogating McKenna’s suggestion regarding the similarity between holding someone morally responsible and having a conversation with her because, even if it correctly characterizes one aspect of moral responsibility, it is not responsive to the basic desert sense of praise and blame that Pereboom insists is the conception of moral responsibility that is centrally at issue between incompatibilists and compatibilists. Therefore, even if McKenna did offer up his conversational model of moral responsibility, it would not constitute an adequate reply to Pereboom.22

I believe that a full defense of compatibilism against Pereboom’s argument is possible if we pay close attention to our practices of holding someone morally responsible. In the remainder of this dissertation, I do just that: offer an account of our moral responsibility practices that leaves no doubt that Plum is morally responsible for killing White. I believe holding a person responsible has two distinct facets, one corresponding to a group of “reactive” emotions, the other to a group of “reactive” nonemotional attitudes. I’ll begin that project in Chapter 2 by discussing reactive emotions. Reactive nonemotions are the topic of Chapters 3 and 4. Chapter 5 concerns one aspect of our practice of holding people accountable: punishment.

1.8 SUMMARY OF CHAPTER 2: REACTIVE EMOTIONS AND PLUM’S MORAL RESPONSIBILITY

In this chapter, I articulate and defend an account of being morally responsible in terms of reactive emotions and then apply this account to Professor Plum to demonstrate

---

22 Some of McKenna’s comments, especially those regarding what he terms the fittingness of certain communications, resonate with the view I defend in the next chapter. But calling blame a conversational “reply” to a wrong brings in conditions of appropriateness for blame that are not what I would term considerations of blame’s “fit.” More on the fittingness of blame in the next chapter.
that he is morally responsible for killing Ms. White. In order to support this claim, I present and defend a broadly Strawsonian account of our moral responsibility practice which understands a person’s moral responsibility in terms of the appropriateness of various reactive emotions we might have toward her. While others have followed Strawson in thinking of moral responsibility in this way, they have not always been careful to distinguish between different evaluations we make of an emotion. Emotions may be appraised as inexpedient, morally wrong or, most importantly for our purposes, fitting. An emotion is fitting with respect to its target just in case it accurately describes that target, a relation analogous to that between a true belief and the world. It is this sense of appropriateness that is relevant to understanding moral responsibility. What theorists impressed by the manipulation argument have missed is that while it might be wrong for us to feel reactive attitudes toward people like Plum who have been manipulated, or such agents may not deserve reactive attitudes from us, reactive attitudes are still fitting for us to feel toward manipulated agents, so Plum is morally responsible on the reactive emotion theory of moral responsibility.

1.9 Summary of Chapter 3: Trust

The account of holding people morally responsible in terms of the reactive emotions that I present in Chapter 2 goes some distance toward capturing the nature of moral responsibility attributions. However, it is not the whole picture. To completely fill out the account of moral responsibility attributions, we need to understand more about nonemotional reactive attitudes. This chapter takes one such attitude, trust, as its topic. I articulate and defend a conception of trust that solves what I call “the trickster
problem.” The problem results from the fact that most accounts of trust treat it similar to, or identical with, relying on someone’s good will. But a trickster could rely on your good will to get you to go along with his scheme, without trusting you to do so. Recent philosophical accounts of trust aim to characterize what it is for one person to trust another so as to avoid this problem, but no extant account successfully does so.

I argue that connecting trust to important, normatively defined relationships like friendship, romantic partnerships and parenting shows us something important about trust. The clearest cases of trust are found within the confines of normatively defined relationships like these, suggesting that there is a normative element to trust. Trusting someone involves not just believing that another person’s good will covers your interactions. Trusting involves believing that, at least in a certain domain of interaction, you are entitled to rely on that person’s good will. This account solves the trickster problem, because a trickster is not entitled to his victim’s good will.

1.10 SUMMARY OF CHAPTER 4: REACTIVE NONEMOTIONS AND MORAL RESPONSIBILITY

In this chapter, I want to use the account of trust developed in Chapter 3 to address a concern about T.M. Scanlon’s account of holding people morally responsible in terms of nonemotional reactive attitudes. First, however, I make interpretive space for the idea that there are some reactive attitudes that are not emotions by briefly returning to Strawson’s “Freedom and Resentment.” I then briefly lay out Scanlon’s theory of holding people morally responsible in terms of nonemotional attitudes like expectations and intentions. Scanlon’s account takes holding someone morally responsible to involve recognizing certain facts about the meaning of her actions for the relationships that you
have with her. While Scanlon’s account does a good job of explaining our practice of holding people morally responsible with whom we already have relationships, it is unclear that it is able to handle cases that involve holding strangers morally responsible.

I argue that the account of trust offered in the previous chapter can helpfully aid Scanlon’s account, in that we generally do have an attitude toward strangers that is appropriately modified by their morally responsible actions—an attitude I call cautious optimism—where we hope that others will act toward us with good will, but do not yet believe that they will. When a stranger acts towards us with ill will, that supports distrusting the stranger rather than continuing to hold the attitude of cautious optimism toward him. Therefore, there is a clear sense in which a stranger’s ill will does properly modify already existing attitudes toward him.

Finally, I turn this account of holding someone responsible in terms of nonemotional reactive attitudes to Plum’s case and show that Plum’s killing of White makes distrust fitting for him. Given that it does, compatibilists should be confident that Plum is morally responsible as he is morally responsible whether or not we understand moral responsibility in terms of emotions.

1.11 Summary of Chapter 5: Punishment

In this chapter, I take up the relation of Plum’s moral responsibility to the institution of criminal punishment. One main locus of the worry about manipulation, I believe, is that if a compatibilist account of moral responsibility is adopted, it might be thought proper to punish a covertly manipulated person who commits a crime and this punishment would be unfair. To helpfully simplify the discussion, I focus on the
justifiability of socially isolating offenders in the state of nature and argue that two nonconsequentialist justifications of socially isolating offenders show us that socially isolating Plum is not unfair. I first lay out a novel justification of socially isolating offenders based on the fact that the offender has shown herself to be untrustworthy. I then show that this account can be helpfully supplemented by Daniel Farrell’s work on the justification of deterrent harms. I apply the distrust and deterrence theories to Professor Plum’s case, showing that on these views he may justifiably be socially isolated. I then show that even though this is a nonretributivist account of the justification of punishment, it can explain why it is essential to punishment that it condemn an offender’s action. To show this, I lay out and extend David Miller and Joel Feinberg’s work on

---

23 I briefly discuss one purely consequentialist justification for social isolation, that defended by Ferdinand Schoeman and Derk Pereboom, and find it wanting. “On Incapacitating the Dangerous,” American Philosophical Quarterly 16, no. 1 (January 1979): 27-35; Living Without Free Will.

I do not discuss retributive theories of the justification of punishment and social isolation because I believe such theories have not been plausibly defended. One of the best sources for objections to such theories is David Dolinko’s work. “Some Thoughts About Retributivism,” Ethics 101, no. 3 (April 1991): 537-559; “Three Mistakes of Retributivism,” UCLA Law Review 39 (1991): 1623.


desert, showing that when we socially isolate offenders in accord with what the distrust and deterrence theories recommend, the offenders deserve that treatment because it expresses anger and distrust toward them. Therefore, it is reasonable to conclude that social isolation done as the distrust and deterrence views recommend would be punishment if we moved it into political society. With little modification, I believe this theory can be applied to punishing Plum in a political society. If so, I have shown that compatibilism is immune to the threat from cases of globally manipulated agents.

---


2.1 INTRODUCTION

In this chapter, I articulate and defend an account of moral responsibility in terms of the fittingness of reactive emotions and then apply this account to Professor Plum to demonstrate that he is morally responsible for killing Ms. White. My argument proceeds as follows:

First, I show that while there are several senses in which an emotion can be evaluated as appropriate or inappropriate, the correct kind of appropriateness for determining someone’s moral responsibility is what I term fittingness. This is the sense in which an emotion is accurate about its target, or gets its target right. To determine when an emotion is fitting, in the relevant sense, requires attention to an emotion’s evaluative presentation: the way the emotion presents the world as being.

In order to focus on an emotion’s evaluative presentation, we need to know which emotions are important for determining someone’s moral responsibility. I argue that anger is the relevant reactive emotion for determining a person’s moral responsibility for actions that are wrong. If a person is the fitting target of anger, she is morally responsible.

---

28 People can also be morally responsible for taking actions that are right or good. A full reactive emotion account would spend more time on positive reactive emotions, like gratitude, whose conditions of appropriateness tracks people’s moral responsibility for positive actions. People can also be morally responsible for performing actions that are
I then present and defend an account of anger’s evaluative presentation and show that this evaluative presentation makes sense of the fact that fitting anger tracks moral responsibility. I show that anger’s evaluative presentation is fitting to have toward Professor Plum and that, therefore, Plum is morally responsible for killing White.

Finally, I offer an explanation of the persistent intuition that Plum is not morally responsible for killing White. Many people have the intuition that anger or resentment is inappropriate toward Plum, but that intuition’s source is ambiguous. I argue that the intuition is better supported by the claim that Plum does not deserve anger than that anger is unfitting toward him. However, I show that whether a person deserves an emotion is irrelevant to that person’s moral responsibility. Therefore, while Plum may well be undeserving of our anger, we cannot conclude from that fact that he is not morally responsible.

2.2 Strawson’s Attitudes

Recall that Strawsonians think of claims about someone’s moral responsibility as claims about the appropriateness of what Strawson terms the “reactive attitudes.” Strawson never provides more than an ostensive definition of those attitudes as including responses like “gratitude, resentment, forgiveness, love, and hurt feelings,” but we can infer some interpretive guidelines even from this list.

---

moranly neutral, just in case they satisfy agential conditions sufficient to make them morally responsible for the action, were it to have been a positive or negative action.

I focus primarily on anger and other negative reactive emotions because I am primarily interested in showing that Plum is morally responsible for killing White.

29 Strawson, “Freedom and Resentment.”
For one, all the attitudes that Strawson mentions are emotions. While there is debate about what, exactly, emotions are, two points help shape our focus here. First, our concern is with occurrent bouts of emotion rather than dispositions to feel emotions. When we say that we resent someone although we do not feel anything toward her at this moment, our claim is most plausibly interpreted as a disposition to have certain negative feelings toward her, act toward her in certain ways, and so on. This makes perfect sense, but it is not our target. Seeking to understand moral responsibility via the emotions involves attention to the conditions that make occurrent bouts of the reactive emotions appropriate.

Second, if we are to make sense of the idea that the reactive emotions can be appropriate or inappropriate in the sense relevant for attributions of moral responsibility, we must assume that these emotions have objects. If we were speaking of mere feelings or moods, which need not be understood as being about the world in any way, then talk of the appropriateness of these attitudes would be strained. Both of the postulates are on display in the following passage from Strawson, where his concern is clearly with a particular bout of emotion and whether it is appropriate for its object:

If someone’s actions help me to some benefit I desire, then I am benefited in any case; but if he intended them so to benefit me because of his general goodwill towards me, I shall reasonably feel a gratitude which I should not feel at all if the benefit was an incidental consequence, unintended or even regretted by him, of some plan of action with a different aim.30

It is generally assumed by theorists working on moral responsibility that figuring out the conditions under which the reactive emotions are appropriate will make clear the

30 Ibid., 63. On the same page, see also Strawson’s discussion of when resentment would be appropriate.
conditions under which it makes sense to hold someone morally responsible. However, there has been insufficient attention to just what sense of emotional ‘appropriateness’ is relevant to the debate. Pereboom prefers to speak of a basic sense of desert, R.J. Wallace invokes desert and reasonableness, while John Martin Fischer and Mark Ravizza simply speak of emotional appropriateness.\(^{31}\) Failing to critically analyze the notion of appropriateness in question is problematic because there are several senses in which we can endorse emotions or find them to be appropriate.

To see some of the plurality of distinct evaluations we can make of an emotional episode, consider these different evaluations of an episode of gratitude.\(^{32}\) On a given occasion, your gratitude toward someone who does something to benefit you might be expedient because it is a pleasurable feeling, morally permitted because it is not wrong to feel happy when someone helps me, or even what to feel when all relevant considerations have been canvassed. But suppose that the person who benefited you is a close rival of one of your friends, Darryl. Suppose that Darryl is very sensitive about his relationship with you and he observes his rival benefitting you. If he notices you feeling grateful about the rival’s action, he might be prone to an envious response, which is more than you feel like dealing with. In such a case, reasons of prudence speak against feeling grateful.

Alternatively, suppose that the benefit that your rival provided was something that Darryl had promised you, but a recent illness prevented him from taking the necessary


action. Feeling grateful toward Darryl’s rival, here, might be thought to be problematic insofar as it might cause Darryl pain when he is already suffering and the reason he was unable to benefit you was not his fault. Seeing your gratitude directed at his rival would just be one more insult on top of his current illness. Moral considerations like fairness, as well as considerations having to do with proper friendship, speak against feeling gratitude in this case.

It might be objected that you don’t have a reason not to feel grateful in the previous example. Instead, because our behavior is generally more under our control than our feelings, you only have reason not to show Darryl that you are grateful to his rival. While it is true that behavior is typically more under our control than our feelings, there is often a very tight connection between feeling and expressing an emotion—especially in terms of involuntary behavior like smiling or blushing. Given the very tight connection between feeling an emotion and expressing it, many people will simply not be able to stifle behavioral expressions of their feelings without suppressing the feelings themselves. To the extent that you are not completely successful at stifling the normal behavioral expressions of your feelings and Darryl is adept at picking up on your current

33 Psychologists have found that the ‘Duchenne smile,’ which is characterized by the activity of both the zygomatic and orbicularis oculi muscles (contraction of both the mouth and eye muscles), is associated with genuine pleasure or enjoyment. Non-Duchenne smiles lack the activity of the orbicularis oculi (eye) muscles and are associated with lack of genuine positive feelings. Paul Ekman, Richard J. Davidson, and Wallace V. Friesen, “The Duchenne Smile: Emotional Expression and Brain Physiology II,” *Journal of Personality and Social Psychology* 58, no. 2 (1990): 342–353.

If, in response to the benefit Darryl’s rival bestows, you exhibit a Duchenne smile that Darryl detects, he will reasonably draw the conclusion that you feel grateful toward the rival.
emotional state, the fact that Darryl would notice whether you show gratitude provides you with a reason not to feel it.

Whether an episode of gratitude is morally permitted, prudentially warranted or what we have reason to feel, all things considered, is a perfectly coherent evaluation, but doesn’t get to the heart of Strawson’s meaning in the above passage. Since Strawson fails to advert to any possible consequences of feeling gratitude, it is unlikely that he means to call gratitude appropriate in his sense in virtue of its being expedient for me to feel toward the person who benefits me. Further, it is doubtful that Strawson means to claim that it is morally permissible for me to feel gratitude toward my benefactor. And without explicitly considering issues like the expediency or moral propriety of the emotion, it’s doubtful that Strawson means to claim that my gratitude toward my benefactor is what I have reason to feel, all things considered. Generic talk of emotions being appropriate or deserved threatens to obscure these different evaluations, at best, or mistake one for another, at worst. In order to adequately develop and defend a Strawsonian account of moral responsibility, these claims must be kept distinct.

2.3 Emotional Fittingness

We can make better sense of Strawson’s previous passage and make progress understanding the conditions under which people are morally responsible by appealing to a convergence in the recent philosophy of emotion literature about an important feature of emotions: the fact that emotions—in Justin D’Arms and Dan Jacobson’s terminology—present the world to us as having “certain value-laden features” or, put
another way, as having specific *evaluative presentations*.\(^{34}\) That is, emotions seem to involve something like perceptions or representations of the world as containing properties like the shameful, the fearsome, the enviable, etc.. To say that emotions involve evaluative presentations is to say that an emotion presents the world to us as containing properties or objects that are evaluated in terms of the above properties. Since reactive emotions are, quite obviously, emotions, the same will be true for them.

Consider gratitude. While we don’t have a ready term for the property invoked by the emotion of gratitude, we can appeal to the same idea by speaking of someone being ‘worthy of’ gratitude. The same will be true of any member of the reactive class. A person, or, more perspicuously, her action(s), can be worthy of resentment, forgiveness, love or hurt feelings.\(^{35}\) When Strawson argues that the same beneficial action, taken without the motive of good will toward me, would fail to be worthy of gratitude, he is implicitly engaged in this kind of property talk.

As D’Arms and Jacobson point out, our practice of arguing about when things have the properties that emotions present them to have—that is, whether things are enviable, worthy of resentment, shameful or worthy of pride—seems to presuppose that we can make sense of a particular kind of emotional appropriateness which is determined


\(^{35}\) The locution “hurt feelings” somewhat strains the point. It would be unusual to say that someone’s action is worthy of our hurt feelings. But when we think that someone hurts our feelings and our hurt feelings get the situation right, we are thinking that hurt feelings are appropriate in a sense that doesn’t involve moral, prudential or all things considered appraisal.
only by the accuracy of an emotion’s evaluative presentation. They term this sense of emotional appropriateness the fittingness of an emotion; fittingness is the sense of appropriateness of an emotion that is analogous to the epistemic relation that holds between a true belief and the world. A belief that is true presents the world as containing certain features or properties that the belief correctly represents the world as containing; a fitting emotion presents the world as containing a particular set of features that the emotion correctly characterizes the world as having.

A return to Strawson’s example of gratitude is instructive. From the previous passage, we can glean Strawson’s view about gratitude’s evaluative presentation, which I think is substantively correct. Gratitude presents the world as containing something roughly like the following: (i) an agent, who (ii) performed an action, which is (iii) motivated by good will toward you, that (iv) benefited you (or someone or something you care about), and is (v) construed as good (for you, or some other person or thing). Should the world contain such a situation, then gratitude’s evaluation is correct and so your feeling gratitude toward the person in question is fitting (the person is worthy of your gratitude). However, should the action not have benefited you, your feeling gratitude toward the person would be unfitting. Similarly, if the person had been

---

36 I am not suggesting that feeling grateful requires you to believe all of these things, or that if you believe all of these things, that suffices for feeling grateful. Rather, I am suggesting something analogous to the experience of seeing a painting hanging on a wall. Your gratitude presents (i)-(v) as being in the world, just as your experience presents the painting as being there. In feeling grateful, you are not required to believe that (i)-(v) hold any more than you are required to believe there is a painting on the wall because you have the relevant experience (suppose you know what you see is produced by a hologram, for example). Similarly, believing that (i)-(v) hold does not suffice for being grateful any more than believing the painting is on the wall means you are currently perceiving that it is (you could have left the room, for example).
motivated by ill will in acting as she did, it would be unfitting for you to feel gratitude toward her. Strawson does not mention these additional possibilities (perhaps because they are too obvious to require comment), but it would also be unfitting for you to feel gratitude toward the person if the beneficial result were not produced by her, if there were no person on the scene at all, or if the action done was in no way good (for you or anyone else). Strawson’s discussion, then, calls our attention to the properties that gratitude is a presumptive response to, and evaluation of. When those properties hold, someone is worthy of our gratitude; gratitude is fitting.

2.4 FITTINGNESS AND MORAL RESPONSIBILITY

Philosophers interested in moral responsibility who think that the key to understanding it will be found in the appropriateness of the reactive emotions should take the sense of appropriateness in question to be that of an emotion’s fit. I’ll advocate for this claim using an argument of attrition to demonstrate why other plausible contenders for characterizing the relevant notion of appropriateness should be rejected, showing fittingness to be the only sense of emotional appropriateness that correctly tracks moral responsibility. Several of the other contenders for characterizing the correct sense of appropriateness have been mentioned already: the moral, prudential or all things considered propriety of an emotion. In this section, I’ll first demonstrate that none of those can be the correct notions of appropriateness and then argue against a more difficult

37 There are many different ways that such an action could be good for me or someone else. It could satisfy a desire, make me feel good, reaffirm a relationship, promote my well-being, etc. Any of these could be enough to satisfy condition (v).
challenge: that the relevant notion of appropriateness is that of an emotion being deserved.

The argument against the first three possibilities is simple. We can see that the sense of appropriateness in question cannot be the moral, prudential or all things considered appropriateness of an emotion by returning to the above example concerning Darryl’s rival. Intuitively, whether or not Darryl’s rival is morally responsible for benefitting you doesn’t depend at all on how Darryl would respond if you feel gratitude toward the rival. On the other hand, whether there is a moral, prudential or all things considered complaint against your gratitude can depend on how Darryl is likely to respond. This entails that general moral, prudential or all things considered considerations do not give us the correct notion of appropriateness for determining someone’s moral responsibility. Further, since the fittingness of gratitude for Darryl’s rival is immune from questions regarding Darryl’s response, the claim that fittingness is the relevant notion of appropriateness for determining the rival’s moral responsibility gets additional intuitive support.

In reply, an objector might urge that while I’m right that an emotion’s general moral, prudential or all things considered appropriateness doesn’t track moral responsibility, there is a specific kind of moral consideration that does track moral responsibility, namely whether or not there is an all things considered complaint against your gratitude in this situation depends on there being other considerations besides fittingness that do not exhaust the considerations relevant to what it makes sense to feel, overall. Someone who was a “fittingness fanatic” might hold that you ought to feel what is fitting to feel, other considerations be damned. While I doubt this is correct, such a person will find herself in agreement with my target conclusion: reasons of fit exhaust the reasons that bear on the evaluation of emotions for the purposes of determining someone’s moral responsibility.

38 Whether or not there is an all things considered complaint against your gratitude in this situation depends on there being other considerations besides fittingness that do not exhaust the considerations relevant to what it makes sense to feel, overall. Someone who was a “fittingness fanatic” might hold that you ought to feel what is fitting to feel, other considerations be damned. While I doubt this is correct, such a person will find herself in agreement with my target conclusion: reasons of fit exhaust the reasons that bear on the evaluation of emotions for the purposes of determining someone’s moral responsibility.
responsibility, namely, whether an emotion is deserved by its target. We can certainly imagine Pereboom pressing this point, for he claims that

For an agent to be morally responsible for an action in the sense at issue in this debate is for it to belong to him in such a way that he would deserve blame if he understood that it was morally wrong, and he would deserve credit or perhaps praise if he understood that it was morally exemplary, supposing that this desert is basic in the sense that the agent would deserve the blame or credit just because he has performed the action, given understanding of its moral status, and not by virtue of consequentialist or contractual considerations.\(^{39}\)

Whether a reactive emotion is deserved is intuitively much more closely connected to whether the person who is the target of that emotion is morally responsible. In spite of this fact, there are cases readily at hand where a person is morally responsible for an action and a reactive emotion is fitting to have toward that person’s action, but the emotion is undeserved. Consider this case:

Ross is a student who has been studying overseas. He decides to return home to the US for graduate school. Without being asked, his parents decide to purchase several expensive items he’ll need, including a TV and a bed.

Intuitively, gratitude is fitting for Ross to feel toward his parents for assisting him in getting ready for graduate school. He needs the TV and bed, and not having to purchase them himself is going to save him both time and money.

But Ross is not grateful to both his parents. He feels very grateful toward his mother, but feels no gratitude at all toward his father. The reason that he does not feel gratitude toward his father is that his father had previously spent many hours “educating” Ross about how men should comport themselves in the world. One lesson Ross’s father tried to drive home is that men should endeavor to feel as little emotion as possible and

---

especially should not feel grateful toward other men. Feeling gratitude demonstrates weakness, according to Ross’s father, since it implies the person feeling gratitude is in need of assistance.

Ross’s mother is distressed at what she views as Ross’s stunted emotional response to his father. She presses Ross about why he feels gratitude toward her, but not his father. She points out that both she and Ross’s father spent several hours acquiring the things for Ross and notes that Ross’s father was solely responsible for purchasing the expensive TV, while she only bought the less valuable bed. “Isn’t it appropriate to feel gratitude toward both of us?” Ross’s mother demands.

In urging Ross that gratitude would be appropriate for him to feel toward his father because it is appropriate to feel toward her, Ross’s mother is implicitly invoking the conditions of gratitude’s fit to make an argument to Ross regarding what he ought to feel. When Ross replies that his feeling gratitude toward his father is inappropriate because his father taught him not to feel grateful toward other men, Ross is invoking the fact that his father does not deserve his gratitude to support his argument about what he ought to feel.

Now consider that while Ross might be correct that his father does not deserve his gratitude, there is no doubt whatsoever that Ross’s father is morally responsible for buying Ross the TV. If anyone is morally responsible for anything, Ross’s mother is morally responsible for buying him the bed and Ross’s father is morally responsible for buying him the TV. So it should be clear that considerations regarding whether Ross’s father deserves his gratitude do not track his father’s moral responsibility, while considerations regarding the fittingness of gratitude for Ross’s father do track his father’s moral responsibility.
moral responsibility. Pereboom is incorrect that the desert of a reactive emotion is the relevant notion of emotional appropriateness for determining someone’s moral responsibility.

Here Pereboom might object that while I have presented a case where someone’s deservingness of a reactive emotion does not track that person’s moral responsibility, I have not presented a case in terms of basic desert, defined as desert where “the agent would deserve the blame or credit just because he has performed the action, given understanding of its moral status, and not by virtue of consequentialist or contractual considerations.” While my example did not appeal to consequentialist or contractual considerations, Pereboom might urge that Ross’s father’s deservingness is not of the relevant basic kind, since whether Ross’s father deserves Ross’s gratitude depends on historical facts about their relationship. This means I am not making claims about Ross’s father being undeserving of Ross’s gratitude “just because Ross’s father bought him the TV.”

I grant that this is true. However, Pereboom is not entitled to argue against my example on these grounds, because it would rule out his own use of the Plum example. Professor Plum has a history with the Neuroscientists, just as Ross’s father has a history with Ross. It is what the Neuroscientists did to Plum in the past that generates the intuition that Plum does not deserve blame, just as what Ross’s father did to Ross in the past drives the intuition that Ross’s father doesn’t deserve Ross’s gratitude. But the claims about Plum and Ross’s father’s desert are not supported solely by considerations that involve simply whether “he performed the action, given understanding of its moral

---

40 Ibid.
status.” So Pereboom cannot object to my example on these grounds, unless he wants to disqualify his own case.

We have now seen a case where the fittingness of an emotion tracks moral responsibility while the deservingness of that emotion does not track moral responsibility. Coupled with the fact that the moral, prudential and all things considered appropriateness of emotions also do not track moral responsibility, this pushes us to conclude that the fittingness of reactive emotions is what philosophers interested in moral responsibility ought to have referenced all along, whether or not they have understood themselves to be doing so. If fact, we can even read Pereboom as implicitly invoking claims about an emotion’s fit with his talk of basic desert—by invoking desert “just because the person performed the action,” perhaps Pereboom was gesturing toward fittingness. Once the distinction between considerations of fit and other kinds of emotional evaluation has been made, we avoid confusion by being explicit that fittingness is the relevant kind of assessment of an emotion for the purposes of determining whether a person is morally responsible.

2.5 Anger and Being Morally Responsible

Supposing that the above argument is correct, we now need to turn to an examination of what features in the world make specific negative reactive emotions—like resentment—appropriate, given that our concern is with Plum’s moral responsibility for doing something wrong. But we now face a problem, which is that there are multiple negative reactive emotions that could be invoked as fitting or not, including blame, resentment and indignation. Progress on understanding the evaluative presentations
necessary for determining the fittingness of the relevant attitudes will be made more
easily if we can focus on just one of these emotions, as we did in the case of gratitude.
But it will be hard to choose, since some theorists prefer to talk about blame, others
resentment, and so on.

I’m not directly interested in the lexical issue about what to call the relevant
attitude, but I need to use some term to pick out the relevant emotion. I prefer ‘anger’ for
this purpose because the emotion I would naturally describe using the term ‘anger’ is
what we might call the “basic root” of blame, resentment and indignation. That is, to
the extent that one is feeling resentful, indignant or blameful, one feels what I would term
anger. I don’t mean to deny that you can coherently describe yourself as blaming
someone (or being indignant at or resenting her) although you don’t feel angry with that
person—you can. But these uses of ‘blame’ fall outside our current area of inquiry,
which is into occurrent reactive emotions and their conditions of fit.

I want to contrast the emotion I am calling ‘anger’ with what I term ‘frustration,’
which I take to be a different, though related, emotion. We sometimes find ourselves to
be emotionally bothered in a phenomenologically similar manner to feeling angry in

41 As we’ve seen above, Pereboom generally speaks of blame.
42 R.J. Wallace prefers to talk of indignation and resentment. Responsibility and the
43 The basic emotions are thought to make up a class from which all other emotions are
derived. There is some dispute about exactly which emotions are basic, but a
representative list might include some or all of the following: amusement, anger,
contempt, contentment, disgust, embarrassment, excitement, fear, guilt, pride in
achievement, relief, sadness, satisfaction, sensory pleasure and shame. This list of 15
editions is from Paul Ekman’s “Basic emotions,” in Handbook of cognition and
emotion, ed. Tim Dalgleish and Mick Power (New York: Wiley, 1999), 45-60. Ekman
defended a shorter list of six basic emotions in earlier work (happiness, surprise, anger,
sadness, fear, disgust/contempt).
situations where we don’t take anyone’s moral responsibility to be involved; suppose the
car won’t start when you need to get to your job interview, for example. While the way
we feel in a case like this is similar to the way we are bothered when someone
intentionally insults us, frustration is directed at a negative situation and so is different
from anger, which is directed at a person. As my inquiry concerns the conditions under
which people are morally responsible for what they do, my focus in what follows is on
anger.

To this point, I’ve argued that fittingness is the relevant notion of emotional
appropriateness for determining someone’s moral responsibility and that on a reactive
emotion account of moral responsibility, anger is the emotion on which such an account
should focus for determining the moral responsibility of someone who has done
something wrong. On the latter point, even Pereboom, who often speaks of blame,
agrees: “of all the attitudes associated with moral responsibility, it is anger that seems
most closely connected with it.” If the argument to this point is correct, we are entitled
to conclude that on a reactive emotion view of moral responsibility, someone is morally
responsible for doing something wrong if anger befits her action. We now need to turn to
a careful consideration of anger’s evaluative presentation in order to ascertain anger’s
conditions of fit. That will allow us to determine if Plum is befitting of anger and thus
morally responsible.

44 Pereboom, Living Without Free Will, 208.
2.6 Anger’s Conditions of Fit

Articulating an emotion’s evaluative presentation is challenging because while emotions have quasi-representational qualities, they don’t describe the things in the world they represent. This contrasts with other mental states like beliefs. For example, beliefs about mountains include thoughts of peaks and slopes, as well as thoughts of rock or some other material; i.e. they describe actual mountains by being representations that have parts that mirror the parts of the mountain. But emotions exhibit their quasi-perceptual character without describing the situation they represent.

Jesse Prinz provides a helpful example showing that it’s possible for something to have representational character without describing what it represents by invoking the beep from a car-mounted radar detector. The beep of the detector represents the presence of police radar. But the beep cannot be broken into other subbeeps. The beep represents police radar, but it does not describe police radar. Similarly, emotions also represent without describing. Prinz’s example not only helps to emphasize that something can have representational character without describing, it also points us in the direction of a method for determining anger’s evaluative presentation. The radar detector is set up by humans to track the presence of police radar with the beeping sound. We can determine what the beep represents by appealing to the functional role that the designers of the detector created it to occupy. Similarly, one way to get a fix on anger’s evaluative

---

presentation is by considering the functional role of anger (though obviously, anger will not have a designer in the same sense).\textsuperscript{46}

Think of the prototypical ways in which we respond when angry. Our heart rate increases, our brow furrows, our muscles tense. Our attention may become fixated on a person whom we take to be acting wrongly. We become primed to vocalize or express our anger toward that person by speaking sharply or yelling. In extreme cases, we are even motivated to fight or attack.

Now consider: what function is served by us having such a state in our repertoire of responses? In what situations would such behavioral responses make sense? There are two related functions that such a state might serve: detecting and warding off dangerous behavior from another agent and detecting an already completed harm or slight in order that retaliation can warn against future such occurrences. Physiological changes like increased heart rate serve to ready our body for action that might be advantageous to counter a threat. It makes sense to closely monitor another person who may be acting in a way that could counter your interests. A communication that her course of action is unwanted could stop that action, or lead to it not being repeated. And, should she continue, an attack might stop her or, if it is too late to do so, an attack might serve to prevent her from acting in that way again.

I believe that we make best sense of being readily susceptible to the normal behavioral upshots of anger if we suppose that anger represents a norm violation. The relevant norms need not be formulated in explicit rules, nor are the only relevant norms

moral norms—anger can serve its social function as a response to breaches of nonmoral norms as well. When the relevant kind of violation is present, then anger’s associated physiological, expressive and motivational tendencies tend to be beneficial to the agent. Of course, this need not be true in every case. Anger might not be beneficial for me to feel toward a person who is very strong and has a very short fuse, no matter how serious the norm he violates. But cases like these do not impugn the idea that anger has a function, any more than the fact that sweating in an extremely humid environment has little cooling effect and over taxes your body refutes the claim that the function of sweating is to cool your body. Sweating still serves an overall biological function even if it is maladaptive in some instances, just as anger may serve an overall social function even if it sometimes proves misguided.

The function of anger is social; the behavioral expressions of anger involve trying to get people with whom we are angry not to violate norms that rule out improper behavior or, if a violation has already occurred, trying to motivate those with whom we are angry not to violate the norms in the future. Note that to say that anger represents norm violations is not to say that it represents mere transgressions, that is, the performance of an action that is ruled out by a norm. A person’s action that transgresses a relevant norm must be the result of ill will or insufficient concern in order to be what I call a violation of that norm. For example, the norm that requires that we refrain from harming is merely transgressed but not violated when someone, while taking proper care, backs her car over the foot of an unseen stranger. To say that anger represents norm
violations is to say that it represents norm transgressions that are the result of ill will or lack of concern for the things the norm protects. 47

Anger’s physiological expressive and motivational tendencies are more likely to be beneficial when they respond to a putative norm violation. Consider that if you are angry when there is no person around, or if the person you’re angry with didn’t perform an action, or her act doesn’t violate a norm ruling out behavior, or she was actually motivated by good will, then anger’s expressive and motivational tendencies will likely not be adaptive for you. (Though, of course, we can imagine unusual cases where they would be.) Lashing out at falling rocks doesn’t make the rocks act differently—and might involve harm to us. Acts of other people that don’t violate norms for proper conduct are unlikely to negatively impact us, so they can safely be allowed to continue. A person motivated by good will is unlikely to be engaging in an action that will transgress a norm. If her good will might lead to harm because she doesn’t know what she is doing, then it makes more sense to simply correct her misperception than to attack, which might cause her to have ill will toward you. 48

The above functional appeal supports the following characterization of anger’s evaluative presentation:

Anger presents the world (i) as containing an agent (ii) who acts (iii) in a way that transgresses a norm on proper conduct (including, but not limited to, moral norms, though the norm need not be codifiable by a rule) (iv)

47 For example, moral norms serve at least to protect the interests of sentient beings. A person who transgressed a moral norm because she failed to care that her action would harm sentient beings would show insufficient concern in the sense I have in mind.
48 I want to emphasize again that the claim is not that being angry in such situations could never serve a person’s interests, but that the behaviors that anger typically produces are unlikely to have positive effects in such situations.
because she is motivated by ill will or has shown insufficient concern, (v) and construes this as bad.\(^{49}\)

But considering anger’s functional role is not the only way we can fix on anger’s evaluative presentation. Further support for this account of anger’s evaluative presentation can be gained by seeing how it tracks situations in which we tend to find it appropriate to be angry. For example, it makes sense of the following example from Strawson:

If someone treads on my hand accidentally, while trying to help me, the pain may be no less acute than if he treads on it in contemptuous disregard of my existence or with a malevolent wish to injure me. But I shall generally feel in the second case a kind and degree of resentment that I shall not feel in the first.\(^{50}\)

Reading the passage in conjunction with my above account of anger’s evaluative presentation implies that if someone treads on my hand with contemptuous disregard of my existence (lack of required concern) or a wish to injure me (ill will), then anger is fitting to feel toward that person—which clearly matches Strawson’s suggestion that resentment is fitting in such a situation. Further, in both those cases it would intuitively seem that the person in question is morally responsible for treading on my hand. But if my hand was injured in a way that was accidental, where someone was trying to help me out of good will toward me, anger would be unfitting for me to feel toward her. Again, this tracks intuitions about moral responsibility, for if the person was taking all necessary care while trying to help, we would intuitively not find her to be morally responsible for treading on my hand. We can also imagine other permutations of the case where my

\(^{49}\) The same point holds here as did with respect to gratitude’s evaluative presentation. I am not claiming that conditions (i)-(v) must be believed by a person in order for her to be angry, nor am I claiming that believing (i)-(v) is sufficient for a person to be angry. \(^{50}\) Strawson, “Freedom and Resentment,” 63.
hand is hurt by the falling of an unstable piece of wood, violating condition (i), where my hand is hurt by someone staggering to get her balance after being pushed, violating (ii), or where nothing in any way bad happens, violating (v). In these cases, anger would be unfitting for me to feel toward the other person, which again tracks intuitions about whether someone is morally responsible.

Consider these cases as well: You might become angry with a neighbor whom you suspect stole your bike, even though you suspect the wrong person. And you might be angry at a political figure whose hands were tied by unavoidable bureaucratic hoops of which you are unaware. In these cases, the people in question either do not violate a relevant norm or they transgress the norm without showing ill will, insufficient concern or insufficient attention, so in these cases you should reject the anger you feel as unfitting.

My account of anger’s evaluative presentation in terms of conditions (i)-(v) is supported by attention on anger’s functional role, by its ability to make sense of situations in which our anger is appropriate in the sense of fitting and also by its ability to make sense of situations in which our anger is unfitting. I could offer other cases that support the account as well, but at this point it behooves us to consider objections to the proposal.

---

51 The sort of badness in question can be as diverse as the goodness that gratitude tracks. But if nothing at all bad occurs, anger is unfitting.
2.7 Two Objections to the Proposal

Desert rears its head again. I argued above that the appropriateness conditions of reactive emotions for the purposes of determining someone’s moral responsibility involve only the conditions of an emotion’s fit and do not involve considerations having to do with whether the emotion is deserved. But we might imagine an objector responding that while I’ve shown that conditions of fit are the appropriateness conditions on reactive emotions that track people’s moral responsibility, I still need to show that deservingness is not, itself, one of anger’s conditions of fit. That is, I need to show that anger’s evaluative presentation doesn’t also include:

(vi) the person who satisfies (i)-(v) deserves anger.

If anger’s evaluative presentation does include a condition involving desert, then, while I will have shown that fittingness is the only relevant sort of emotional appropriateness for the purposes of determining someone’s moral responsibility, I will not have shown that claims about whether or not someone deserves reactive emotions are irrelevant for determining people’s moral responsibility.

To make the case for the inclusion of (vi), the objector would need to present cases where, without the addition of (vi), the fittingness conditions of anger fail to track people’s moral responsibility for doing wrong. I do not believe cases of that sort can be offered, so let me now offer an argument in support of the opposing claim: if we did include (vi) as a condition on anger’s fittingness, anger would fail to track moral responsibility. If I can show this, then we either give up (vi) as a condition on anger’s fittingness, or we give up the idea that anger is a responsibility-tracking emotion. Since the second possibility violates the Strawsonian methodological assumption of
understanding people’s moral responsibility for wrong actions in terms of the appropriateness of attitudes like anger, we have to give up (vi) as a condition on anger’s fittingness.\textsuperscript{52}

To see why (vi) can’t be a condition on anger’s fittingness, consider this case:

Your new bike is stolen from behind your residence. The police contact you and tell you they have the perpetrator in custody. He is a neighbor whom you know to be impoverished, and the police inform you that his own bike—which he relied on to get to work—was also recently stolen. As you consider the situation, your initial anger at your neighbor changes to sympathy. If anything, you are now angry with the person who stole \textit{his} bike.

For many of us, the emotional arc traced by this example will be familiar. Initial infuriation at wrongdoing can be quickly tempered by the thought that the perpetrator was wronged, himself, and that therefore he doesn’t deserve your anger. But note that as you find your neighbor no longer deserving of your anger, you need not be entertaining any thoughts about how the recent wrong against him and his economic hardship impacted his moral responsibility for the crime or the wrongness of what he did. You could, to be sure, imagine that your neighbor is so beset by material depravity that he literally needed a bike \textit{immediately} in order to preserve his already tenuous situation. If all options other than taking your bike are much worse for your neighbor, you certainly might suppose that he didn’t do anything wrong in taking the bike.

\textsuperscript{52} Another option would be to give up on a Strawsonian approach to moral responsibility. This would be a mistake, as there are good reasons to prefer such an account. I won’t go into them here other than to note that philosophers have struggled to say much about what it is to be morally responsible other to be a fitting target of the reactive attitudes. Further, the Strawsonian account does the best job of giving voice to our intuitive conception of moral responsibility in a way that characterizes someone’s being morally responsible in a sense prior to any overt response like punishment.
But while you could be imagining your neighbor with his back against the wall, you need not—and most of us do not have to—actually imagine that situation in order to find your neighbor undeserving of your anger. Most of us will only have to suppose that your neighbor’s material situation is quite a bit worse than ours to have sympathy for the fact that his own bike was recently stolen. Just supposing that, I submit, will be enough to make our anger at him seem undeserved.

The point of this example is to show that we can find someone who acted wrongly to be undeserving of anger without supposing anything at all about that person’s agency or any other factor that would seemingly impact her moral responsibility for doing wrong. People who are poorly off financially pretty clearly take many actions for which they are morally responsible. If someone applies for and gets a job when she really needs the money, she is morally responsible for getting the job. Further, as the bike case shows, we regularly take people who are in financial need and cope by stealing to be morally responsible for doing so. Material deprivation is not a general moral responsibility-defeater, nor is being the subject of a recent wrong. So the mere fact that your neighbor was materially deprived and was wronged doesn’t impact his moral responsibility for coping with his situation by stealing your bike.

I submit that the reason that we find anger to be undeserved toward someone who wrongs us but is materially much worse off is that we take our anger at him to be undeserved in the sense of unfair. Not only is this person’s life much less pleasant than ours, we imagine, but he now has to also contend with our being angry with him on top of everything else. Being angry with him would just pile one more insult on all his other injuries. But note that if the deservingness of anger toward someone can vary with his
material situation and the fact that he has been wronged, then the deservingness of anger pretty clearly cannot be a condition on moral responsibility, since the mere fact that someone is materially worse off and was recently wronged does not affect the person’s moral responsibility for what he does.

The way to avoid the unintuitive claim that your neighbor lacks moral responsibility for taking your bike, while at the same time offering a theory consistent with the intuition that anger is undeserved in such a case, is to note that your anger fits what your neighbor did, while admitting that at the same time your anger toward him is undeserved in the sense that it is unfair. Anger fits what your neighbor did because he violated a moral norm and he was insufficiently motivated by the fact that his action was wrong. Anger at your neighbor is undeserved, not because we doubt that he is morally responsible for taking the bike, but because your neighbor’s relative deprivation and the wrong he recently suffered makes us doubt the fairness of responding with anger to his actions. The fact that your neighbor is undeserving of anger at the same time as he is morally responsible shows that deserving anger is not a condition on anger’s fittingness.

I have now addressed the objection that deserving anger is a condition on anger’s fittingness, but we can imagine one more objection to the account. This objection suggests that, while desert is not a condition on anger’s fit, there are other conditions on anger’s fit beside the five I presented. Such conditions might include (vi) that the agent’s action was not casually determined, (vi) the agent’s action was not produced by antecedent causes that are out of his control to which his actions are linked by causal laws

---

53 These additional conditions are inspired by worries about how determinism and causation might impact a person’s moral responsibility.
of the kind that govern other events in the universe, or (vi), the agent lacked alternative possibilities. By proposing any of these conditions, an objector would be pushing back directly at my account. He would be arguing that anger is unfitting when it violates (vi), (vi) or (vi).

My initial reply to this objection emphasizes the earlier account of anger’s evaluative presentation in terms of its function. The four features of anger’s evaluative presentation that I have emphasized make sense given anger’s role as a norm violation detector. One way to offer a positive case for the inclusion of a condition like (vi), (vi) or (vi) would be for the objector to demonstrate that the addition of a condition like (vi), (vi) or (vi) is necessary to make sense of anger’s role in our mental life as a norm violation detector. But such a case is not available, for conditions like (vi), (vi) or (vi) do not concern whether an agent has violated a norm, but how an action that violates a norm was produced. Anger would not function more successfully as a norm violation detector if its evaluative presentation included any of these conditions.

Perhaps, instead, a case can be made for including (vi), (vi) or (vi) by reflecting on common and intuitively compelling cases. After all, another consideration I claim supports my account of anger’s evaluative presentation is that it makes sense of a wide variety of cases where we take anger to be appropriate. So if an objector could argue for (vi), (vi) or (vi) on similar grounds, she would have a good case for such a condition being a part of anger’s evaluative presentation.

I am similarly skeptical about the success of this argument, however. The most compelling sort of case that could be offered for the inclusion of (vi), (vi) or (vi) would
be a manipulation case like that of Professor Plum. But that case has a defect that prevents it from offering a compelling reason for the inclusion of (vi), (vi), or (vi), namely, the presence of the nefarious neuroscientists. To see why this is a defect, let’s consider a general lesson from the cases involving Ross and his father and your neighbor and your bike.

In Ross’s case, we saw that his father’s past actions functioned to present a moral objection to him receiving a positive response from Ross for what he did. In your neighbor’s case, we saw that your neighbor’s material situation, along with the fact that he recently suffered a wrong, functioned to present a moral objection to responding to his wrongdoing with fitting anger. But the role played by the moral objection did not bear on the moral responsibility of Ross’s father or your neighbor. The neuroscientists play a similar role in Plum’s case: their actions affect the fairness of our responding to Plum with anger without impacting the fittingness of anger toward him. Consider Plum’s likely response if we were to be angry with him for killing White. The most obvious thing for him to say is something like this: “Look, don’t be angry with me...if the neuroscientists hadn’t done what they did to me, I wouldn’t have killed Ms. White. They’re the ones who are really did something wrong. Be angry with them, not me.”

54 Saul Smilansky lists manipulation cases as one of three issues that push him toward skepticism about free will and moral responsibility (and so, perhaps, to preferring (vi), (vi), or (vi) as part of anger’s evaluative presentation). He writes, “The second challenge to compatibilism is more internal, and begins with our intuitions on specific cases (such as instances of gross manipulation). Those cases, which we all see to be tainted, are then shown, through minute modifications, not to be fundamentally different from the cases that the compatibilist takes to be models of control and responsibility.” “Fischer’s Way,” 151.

55 Plum might well also urge that it is the neuroscientists, not him, who should be punished for killing Ms. White. I take up this issue in Chapter 5.
Note that Plum saying that the neuroscientists are the ones who “really did something wrong” implies that we shouldn’t think of him as having done anything wrong. But we can reject this aspect of Plum’s appeal, for he did do something wrong: he killed White. What seems correct about Plum’s appeal is something else that is implied by it: we have more reason to be angry with the neuroscientists than we do with him. After all, the neuroscientists not only made Plum kill White, they seriously wronged Plum by mentally manipulating him. Plum’s plea against our anger, then, does not call the fittingness of our anger at him into question. Instead, it calls into question whether Plum deserves our anger when there is another party, the neuroscientists, who wronged him and seem to deserve our anger even more than Plum does. As Pereboom puts it, we should have sympathy for Plum’s situation. But a sympathetic response to Plum only shows us something about moral responsibility if we can be sure our sympathetic responses are reactions to some moral responsibility-relevant feature of the case. I have argued that this is not the case. Our sympathetic responses to Plum are produced by facts about his case that attenuate his deservingness of our anger, and someone’s deservingness of anger is not the sort of emotional appropriateness that tracks moral responsibility. Therefore, Plum’s case does not give us any reason to add a condition like (vi)$_1$, (vi)$_2$ or (vi)$_3$ as part of anger’s evaluative presentation.

56 Again, Pereboom’s case is meant to preserve all important aspects of Plum’s agency.
57 “A Hard-line Reply to the Multiple-Case Manipulation Argument,” 167.
58 Of course, an incompatibilist might find some other motivation for adding a condition like (v)$_1$, (v)$_2$ or (v)$_3$ to the set of anger’s fittingness conditions. To the extent that I haven’t canvassed all possible motivations, my argument is incomplete. However, as I view manipulation cases to be the best motivation for adding such a condition, I’m content to have argued that they don’t support adding it.
2.9 Plum’s Moral Responsibility

While the conclusion of this section has already been prefaced, we now need to turn to an explicit consideration of what the account of anger’s evaluative presentation tells us about Professor Plum’s moral responsibility for killing Ms White. In order to determine whether anger is fitting to feel toward Plum, we should see if the four conditions that characterize anger’s evaluative presentation are met. Recall that anger’s evaluative presentation involves:

Anger presents the world (i) as containing an agent (ii) who acts (iii) in a way that transgresses a norm on proper conduct (including, but not limited to, moral norms, though the norm need not be codifiable by a rule) (iv) because she is motivated by ill will or has shown insufficient concern, (v) and construes this as bad.

Conditions (i) and (ii) are clearly satisfied. Plum is a person who has committed an act. Condition (iv) is also obviously met, for Plum’s action is generated by a process of reasoning that specifically privileges his own desires to kill White for some small personal benefit over considerations concerning the harm to her. Plum at least lacks good will and is likely demonstrating ill will. Finally, conditions (iii) and (v) are satisfied because Plum transgresses the norm against killing innocent persons. Since anger is fitting to feel toward him, Professor Plum is morally responsible for killing Ms. White.

2.10 Conclusion and Error Theory

In this chapter, I have argued that the fittingness of a reactive emotion, not whether it is deserved, is the relevant notion of appropriateness for determining when someone is morally responsible. Further, I have argued that anger is the relevant reactive emotion for determining someone’s moral responsibility for actions that are wrong. On
those grounds, I concluded that for someone to be morally responsible for performing a wrong action is for anger to befit the person who committed it. Then, I presented an account of anger’s evaluative presentation that enabled us to see that anger is fitting in response to someone just in case that person acts in a way that violates a norm on proper conduct. I defended that account of anger’s evaluative presentation against objections that it erroneously left off an additional condition concerning desert or a condition concerning the historical production of the action. I showed that anger is fitting for Plum, so I conclude that a reactive emotion account of moral responsibility requires us to say that Professor Plum is morally responsible for killing Ms White.

I do not expect all readers to agree with me, however. The intuition that anger is inappropriate for Plum and that he therefore cannot be morally responsible for killing White is persistent for some readers, so let me explain the source of that intuition in a way that shows the intuition to be consistent with Plum being morally responsible for killing White.

D’Arms and Jacobson point out that often “one is reluctant to endorse, in any respect, a feeling one deems to be wrong; and that to judge an emotion to be fitting is to endorse it—albeit only in a particular respect.”59 Consider, for example, the awkward position we are placed in if a colleague tells an offensive joke. Laughing at the joke seems to endorse the offensive view expressed by the joke, so a common response is to deny that the joke is funny at all. But surely some offensive jokes are funny—that is why they are so pernicious. Unfunny, offensive jokes are boring and tend not to be repeated. As such, they have little power to support the offensive view evinced in them.

Compatibilists should thus emphasize that the likely source for the fact that many people find anger inappropriate for Plum is a similar phenomenon to when we are confronted with an offensive joke that, while funny, is also reflective of a morally odious view. Our amusement at such a joke is stifled by our reaction to the obnoxious view contained in it. Similarly, our fitting anger at someone who commits a grievous wrong is blunted in Plum’s case by our sympathetic response to the serious wrong done to him, which generates the intuition that anger at Plum would be undeserved or unfair. To infer that because anger is inappropriate to have toward Plum on these grounds, anger is not fitting toward him commits what D’Arms and Jacobson call the moralistic fallacy: the tendency to infer, simply from the fact that it would be wrong to feel an emotion, that the emotion does not fit its object.\(^6^0\) We can expect, then, that in situations where a person has been wronged or is otherwise poorly off, we will be reluctant to endorse negative feelings toward that person—it seems like piling insult on injury. This natural reaction may well be laudatory, but if this is the motivation adduced to argue that Plum is not morally responsible, compatibilists should resist, for they now have their own explanation of people’s reactions to cases like that of Professor Plum. The intuition that anger is inappropriate for Plum is produced by the unfairness of being angry with him, not by anger’s failure to fit what he did.

Plum may be undeserving of our anger, in a moral sense. The wrongness of feeling anger toward him might even show that, all things considered, we shouldn’t feel anger toward him. But we cannot conclude from either of those claims that it is not

\(^{60}\) D’Arms and Jacobson list an impressive catalog of philosophers who commit the fallacy. See Ibid., 76-86.
fitting to feel anger toward him. After all, as these claims might be true of your bike-stealing neighbor who is clearly morally responsible for stealing your bike, we cannot conclude that Plum is not morally responsible because of a moral objection against being angry with him. Thus, we cannot infer, just from the fact that we feel anger at Plum to be inappropriate, that he is not morally responsible.

There is a certain irony here linked to Strawson’s arguments in “Freedom and Resentment.” While attempting to find common ground between compatibilists who appealed to the efficacy of punishing offenders and incompatibilists who worried about the fairness of doing so, Strawson chided compatibilists for failing to notice that we have a concept of moral responsibility in terms of the appropriateness of the reactive emotions that precedes any overt behavioral reaction to people and that, on this conception of moral responsibility, the efficacy of such attitudes is the wrong sort of condition to offer on these emotions’ appropriateness. In invoking the efficacy of overt behavioral reactions like punishment, or the efficacy of the reactive emotions, optimists were changing the subject.

My account exposes a similar error in the arguments of some incompatibilists. Rather than focus on the conditions of fit of a reactive attitude, some incompatibilists have been guilty of surreptitiously appealing to other conditions of appropriateness. Considering whether a reactive emotion is deserved is common, but amounts to changing the subject in a debate about someone’s moral responsibility.

---

61 This is part of Strawson’s concern in section II. “Freedom and Resentment,” 60-62. Note this suggestive quotation: “Now my optimist, being the sort of man he is, is not likely to invoke an intuition of fittingness at this point. So he really has no more to say.” [Emphasis added.]
3.1 INTRODUCTION

In the previous chapter, we saw that on my reactive emotion account of moral responsibility, Professor Plum is morally responsible for killing Ms. White. But the reactive emotion view, while correctly tracking Plum’s moral responsibility, does not provide a full account of the attitudes implicated in our moral responsibility practices. A complete account of such attitudes will require characterizing nonemotional reactive attitudes. In the next chapter, we’ll meet a general account of holding someone morally responsible in terms of nonemotional reactive attitudes, but we’ll also see that account has a problem explaining in what sense we hold strangers morally responsible. To solve that problem, I’ll need to invoke the attitude of trust. In order to helpfully do so, we first need to understand trust in some detail. So, for the moment, this chapter sets aside Plum’s case to focus on developing a conception of trust that is immune to objections that previous accounts of trust have faced.

The account of trust I articulate and defend in this chapter aims to solve what I call “the trickster problem.” The problem results from the fact that most accounts of trust treat it similar to, or identical with, relying on someone’s good will. But a trickster like

---

Bernie Madoff could rely on your good will to get you to go along with his scheme, without trusting you to do so. Therefore, any account of trust that treats trust as relying on another’s good will is subject to the trickster problem. Recent philosophical accounts of trust aim to characterize what it is for one person to trust another so as to avoid this problem, but no extant account successfully does so. One reason for this is that while theorists interested in trust have standardly used the phenomenon of feeling betrayed as a test for when trust is present—because it is assumed that when your trust is violated, feelings of betrayal will be warranted—there has been almost no attention paid to the phenomenon of betrayal. So while it is generally agreed that feeling betrayed is fundamentally different from feeling let down or disappointed, there has been little attention to the distinction itself. Getting clear about the distinction between feelings of betrayal and feelings of disappointment will point us in the direction of an account of trust that can successfully deal with the trickster problem.

I’ll begin the chapter by laying out some generally accepted ideas about trust—its relation to the ideas of reliance and good will and also the idea that trust that is violated licenses feelings of betrayal. I’ll then introduce Karen Jones’ account of trust, which holds that trust is “an attitude of optimism that the good will and competence of another will extend to cover the domain of our interaction with her, together with the expectation that the one trusted will be directly and favorably moved by the thought that we are

---

However, I’ll demonstrate that Jones’ account is importantly flawed in identifying trust with an affective attitude, specifically, hope. Jones’ attempt to analogize trust to hope makes the account treat people as trusting others when there is clearly no trust. I show that trusting someone is better understood as believing that the good will and competence of that person will be operative in our interactions.

In the second part of the chapter, I use the trickster problem to show why even this improved account of trust cannot say, correctly, that Madoff does not trust his victim to do as he wants. I then turn to an explicit attempt in the literature to solve the trickster problem. Pamela Hieronymi expands on Richard Holton’s account of trust as relying on someone from the participant stance to deal with the trickster worry. However, given plausible assumptions about what the participant stance involves, this idea does not solve the problem.

Next, I examine the phenomenon of feeling betrayed more closely and develop an account of feelings of betrayal that understands the unique character of feeling betrayed in terms of feeling that an important relationship has been ruptured or violated. If this is so, then it makes sense to connect the account of trust on offer with important, normatively defined relationships like friendship, romantic partnerships and parenting. The clearest cases of trust are found within the confines of relationships like these, suggesting that there is a normative element to trust. Trusting someone involves not just believing that another person’s good will covers your interactions; trusting involves believing that, at least in a certain domain of interaction, you are entitled to rely on that

64 “Trust as an Affective Attitude.” Ethics, Vol. 107, No. 1, (Oct., 1996), pp. 4-25. p. 4
65 “The reasons of trust.”
66 “Deciding to trust, coming to believe.”
person’s good will, that the other party owes you good will. This account solves the trickster problem, because a trickster like Bernie Madoff is not entitled to his victim’s good will. It is his realization of this fact that explains both why Madoff does not trust his victim to go along with his scheme and why it would be inappropriate for him to feel betrayed if his victim does not do as he wishes.

3.2 Trusting: The Basic Model

Let me say a bit more about the area I mean to investigate, which is complicated by the fact that the word ‘trust’ is used on different occasions to mean very different things. English allows both of the following usages: "We trust you have enjoyed your flight with Air New Zealand," and "Othello's trust in Iago was misplaced." In the first instance, ‘trust’ is being used to mean something like ‘expect’ or ‘hope.’ In the second, ‘trust’ picks out interpersonal trust, which is my target in this paper. I won’t, then, be directly concerned to account for usages of ‘trust’ that deviate from the interpersonal phenomenon. Examples of the kind of trust I have in mind include the trust that Bernie Madoff’s investors had in him as their friend and financial advisor, and the trust that Elizabeth Edwards previously had toward her husband John, the former candidate for the Democratic Party’s nomination for the US Presidency.

As a first gloss, then, what is it for one person to trust another in the way that

---

Madoff’s clients trusted him and Elizabeth Edwards trusted John? Philosophers interested in trust have standardly answered this question by following Annette Baier in grafting their accounts of trust onto a characterization of another attitude: reliance. 69 Reliance is a mental attitude characterized by an expectation that something will happen. As such, it is an attitude that we can have toward other human beings, but also toward inanimate objects, plants, animals, natural events, etc. You rely on the grocery store to have food stocked and the bus to be one time (or, at least, mostly on time). You also rely on your computer to start up when you push the power button. Insofar as Madoff’s clients trusted him, they relied on him, and insofar as Elizabeth once trusted John, she relied on him as well.

Trusting someone, however, is thought to involve more than simply relying on that person. Baier notes that “The comedian, the advertiser, the blackmailer, the kidnapper-extortioner, and the terrorist” 70 all rely on the other people with whom they interact to do certain things, but cannot be said to trust them. Part of our reason for saying that these people do not trust those they rely on has to do with what is at stake for them if their reliance is disappointed. They may be annoyed, angry, or perhaps surprised in a similar manner to how you might feel if your computer did not turn on when you wanted to use it. In all these cases, the person in question feels what we would describe as disappointment, or perhaps annoyance, but does not feel betrayed, as would be appropriate if a trust had been violated. So it is generally assumed that an important test

69 Annette Baier first connected and distinguished trust and reliance in her “Trust and Antitrust,” which began serious inquiry into trust as philosophical project. Baier, “Trust and Antitrust.”
70 Ibid., 234-5.
for the presence of trust in a given situation is whether a person would reasonably feel betrayed in that situation.

Why can trust be betrayed, but not reliance? The canonical answer, again due to Baier, is that trust involves relying on the good will of another person, as opposed to relying on that person’s “dependable habits,” where these habits could involve ill will or other kinds of motives distinct from good will. An advertiser may expect to see sales for a product increase as the result of the typical formula of putting young, attractive people in the product shoot, but a customer does not betray her if the customer does not respond to her ad in the expected way. This is because the advertiser’s reliance on the customer’s reaction to the product pitch does not depend on the customer’s good will, but on a dependable transference of the customer’s attraction toward the models onto the product.

The idea that trust involves relying on someone’s good will seems intuitive and has been adopted by most theorists of trust. Trust is also thought to involve assumptions about the other person’s competence. As Karen Jones notes, “some people have very good wills but very little competence, and the incompetent deserve our trust almost as little as the malicious.” It is generally agreed, then, that to trust someone involves relying on her good will and competence.

3.3 Jones’ Improvement

Baier’s account of trust as reliance on a person’s good will doesn’t get us far enough, however. General moral claims require that people bear us some level of good will. If we anticipate another person will act as morality demands, we might reasonably

rely on her good will. But trusting is more selective. Trust requires a more personal connection with a person than simply expecting that that person’s good will covers your interactions. Trusting someone also involves assumptions about that person’s motivations—their reasons for acting toward you with good will and what that implies about the circumstances in which you can expect it.

Karen Jones’ account of trust captures this idea. Jones understands trust as

[A]n attitude of optimism that the good will and competence of another will extend to cover the domain of our interaction with her, together with the expectation that the one trusted will be directly and favorably moved by the thought that we are counting on her.72

The idea that the person trusted is being directly and favorably moved by the thought that the one trusting is counting on her should be understood as the person trusted giving significant weight to the fact that she is being counted on. It is not required that this motivation be overriding in the sense that it trumps all other considerations—depending on the circumstances, other considerations may carry the day. However, as Jones notes,

one would not trust if one thought that the fact that one was counting on someone, while always being taken into account, would nonetheless be reliably overridden by other considerations. Were that the case, then, from the point of view of the trustee, the other would appear unwilling to give enough weight to the thought that she was counting on her.73

The person doing the trusting expects that the trustee will be moved by the fact that the truster is counting on her and not simply by a generally benevolent motivation toward all people, for example. Jones’ account captures the idea that trusting someone involves

72 Ibid., 4.
73 “Trust as an Affective Attitude.” p. 8.
taking that person to be motivated in a particular manner, namely, by the fact that the one trusting is counting on her.

3.4 **The Affective Problem with Jones’ Proposal**

My points of disagreement with Jones’ account of trust concern the nature of the attitude that is directed at the good will and competence of another person when you trust her and what that implies about when we can sensibly say that one person trusts another. All parties to the debate believe that trust involves an attitude toward the good will and competence of someone else; Jones thinks that trust requires that attitude to be an **affective attitude of optimism**. She believes this because she wants her account of trust to be able to account for the fact that when we trust someone, we have a tendency to interpret that person’s behavior in a positive light. For example, Jones mentions the fact that in Shakespeare’s play, *Othello*, Othello’s trust in Iago makes Othello unable to perceive that Iago is attempting to harm him, even though the signs that Iago presents a danger are obvious for someone not already heavily biased in Iago’s favor.

Jones believes that this tendency to look on the bright side can best be accounted for by an affective attitude, and that in this sense “trusting is more like hoping than like predicting.”\(^{74}\) Jones wants to draw an analogy between her account of trust, which is attempting to capture our tendency to positively interpret others, and an account of the emotions that emphasizes emotional states as being partly constituted by tendencies of interpretation. For example, Jones claims that the object of resentment might be thought to be a “manipulative exploiter.” If this is correct, then if a person already resents her

---

\(^{74}\) Jones, “Trust as an affective attitude,” 15.
neighbor, it will be hard to get that person to attend to evidence that her neighbor is not a manipulative exploiter.

This is insufficient evidence, however, to support the conclusion Jones wants to draw about the similarity between trust and hope. It is true that trusting someone can result in a tendency to discount evidence of behavior that would undermine trust, but there is no need to posit that an attitude like hope is involved in order to generate this psychological result.\textsuperscript{75}

This is because the kind of bias involved in trust—a bias in favor of the person trusted—can be accounted for by the fact that biases in judgment can occur as the result of a person holding either cognitive or affective attitudes.\textsuperscript{76} For example, the halo effect is a cognitive bias that tends to make perceptions of specific traits in an object or person fit previous global evaluations of that same object or person. Finding that a person is likable, in general, tends to make you perceive them as having other attractive qualities. Finding someone to be physically attractive may lead you to perceive him as kind, or finding someone to be courageous may lead you to perceive her as having good judgment.\textsuperscript{77} The halo effect or some other cognitive bias, then, could account for

\textsuperscript{75} I’m not denying that it’s possible for positive affective attitudes to lead us to generally interpret people’s behavior or qualities positively. I’m simply arguing that it’s possible for cognitive attitudes to generate the same result.

\textsuperscript{76} Some of the best known research into cognitive biases is collected in the volume of Daniel Kahneman, Paul Slovic, and Amos Tversky, eds., \textit{Judgment under uncertainty: Heuristics and biases} (Cambridge University Press, 1982).

\textsuperscript{77} The halo effect was first empirically demonstrated in E. L. Thorndike, “A Constant Error in Psychological Ratings,” \textit{Journal of Applied Psychology} 4, no. 1 (1920): 25–29. Thorndike asked commanding officers to rate soldiers that served under them and found that the officers’ ratings showed a high degree of correlation between positive and negative traits, respectively, in each individual soldier.
Othello’s inability to perceive Iago’s machinations without that bias being accounted for by presuming that Othello’s attitude toward Iago is characterized by hope. It is equally plausible that another attitude, like a belief, can also support Othello’s trusting bias.

Jones’ affective proposal is also problematic for analyzing trust because the account leaves open the possibility that you could trust someone when you hope they will do something but do not believe that they will. But this is false. As Hieronymi points out, “to the extent that you lack confidence in a person’s trustworthiness in some matter, to that extent it also seems correct to say that you do not trust that person in that matter.”

To trust someone, you must have more confidence in her than is captured by saying that you hope that person’s good will and competence will govern your interactions.

We can see this point by considering Elizabeth Edwards’ attitudes toward John after his infidelity is revealed. Given the nature of his work, John travels frequently and spends a significant amount of time away from Elizabeth. During those times, Elizabeth has to decide how to react to their newly impaired relationship. Suppose that there are really only two viable options: to try to go on much as before, throwing herself into her own work, or to try to obsessively monitor John’s actions from afar by calling him and his aides, persistently asking for updates on whether he is with any other women.

To the extent that Edwards does not monitor John’s actions and behaviors because

---


The halo effect is regularly invoked today to explain other social phenomena. For example, it is touted as the reason for the reemergence of Apple’s Mac computers, based on the halo effect generated by positive associations with Apple’s iPod MP3 players.  

she believes that his good will and competence will prevent him from being unfaithful, Elizabeth trusts John.\(^{79}\) Suppose that John has come clean about all his past infidelities and guarantees that he will no longer seek out new paramours. If Elizabeth finds that she believes him, she has some reason to trust John. If she believes that John will not waver when the next chance arises because he will be moved by her and the thought of their relationship, then Elizabeth has a reason based on trust not to worry excessively and spend all her time monitoring his behavior while he is away.

But Elizabeth could also decide for other reasons that monitoring John is not the thing to do. She might actually be uncertain about whether or not his entreaties have been genuine and whether, even if they were, he is competent enough to override sexual urges toward other women. Given her uncertainty, she might feel that the only thing to do is to turn him loose and see what happens. If he doesn’t betray her again, she will have evidence that will allow her to trust him more easily. Thinking to herself, “I have to try to trust him if we’re going to get through this,” she might steel herself and try to focus on other things. Or, alternatively, Elizabeth might be focused less on the health of their relationship and more on the fact that if she obsesses about John while he’s gone, she isn’t able to put on a brave face for their children, who are also very upset by their father’s transgressions. She might be hoping that he is able to live up to his commitment to her, without really believing this is the case.

If Elizabeth decides not to monitor John while he is away for any of the latter reasons, she will not have made her decision because she trusts John. Instead, she will

---

\(^{79}\) I here follow a line of argument that Hieronymi deploys against Holton. Ibid., 216-221.
have decided not to monitor him because, given the costs and benefits associated with monitoring, it makes more sense to try to act normally while taking a gamble on John’s future behavior. In this case, Elizabeth entrusts the future success of their relationship to John, but she does not do so because she trusts him.

The lesson here is that it is possible to entrust someone with the care of something, even yourself, for reasons that go beyond the reasons that ground trust. In this sense, entrusting is like acting (that is, literally performing an action) and can be undertaken for all the familiar and myriad reasons that we act. (Think of the connotation of “entrusting,” namely, handing something over to someone else.) Trust, on the other hand, is an attitude, not an action, and the considerations that support a well-grounded attitude of trust diverge from the reasons that would support entrusting. Therefore, it makes sense to attribute trust to someone only when that person believes that the trustee will come through, not when she decides to take a gamble on the trustee’s future behavior.

The idea that trust involves believing in others can be captured by modifying Jones’ proposal as follows:

To trust someone is to believe that the good will and competence of that person will extend to cover the domain of your interaction with her because the one trusted will be directly and favorably moved by the thought that you are counting on her.

This formulation is preferable to Jones’ account for three reasons. First, it gets correct the core attitude of trust. Trusting someone involves believing certain things about that person, not hoping that those things are the case. Second, this formulation emphasizes that when we trust, the reason that we believe the good will and competence of another person will govern her interactions with us is that we take that person to be moved by the
fact that we are counting on her. Jones’ original account decouples the reason that the person trusting is counting on the trustee from the belief that the trustee will respond with good will. It thus allows that the trustee might respond with good will, and that the trustee is moved in the right way, but that the trustee’s good will is not produced by the trustee being moved in the relevant fashion. It is better to emphasize that when we trust, we expect the trustee’s good will to be produced by her realization that we are counting on her.

Third, this formula emphasizes that while it may be true that we often have generally positive attitudes toward those we trust, positive feelings toward someone are not required in order to trust. To see this, consider the attitudes two prizefighters might take toward each other. Suppose they engage in a particularly brutal form of pugilism with minimal rules that requires participants to fight as hard as they can for as long as they can, until one of the parties is knocked unconscious. The last person standing is the victor. Imagine that the participants in this vicious sport trust each other to follow the minimal rules and fight as hard and as ruthlessly as they can to knock each other out, but that in most other respects they hate and despise one another. Suppose they believe that the other fighters are generally untrustworthy, as even being invited to participate in their sport usually requires a campaign of back-stabbing and underhanded dealing. However, they presume that all the prizefighters are jointly committed to upholding the integrity of their sport.

Imagine that in one battle, one of the fighters purposely throws the match. The other feels enraged and betrayed; fittingly so, even though she is judged to have won the battle. The fittingness of feelings of betrayal would be mysterious in this case if trusting
someone required that you have a generally positive attitude toward the other person. Of course, trust is consistent with such positive feelings and it may be that positive feelings toward those you trust in most cases leads you to have an overall positive evaluation of those you trust, but the above case shows that such feelings are not required for one person to count as trusting another.

3.5 The Trickster Problem

The above account of trust is plausible and is, I think, better than any other account of trust on offer. Even so, it is vulnerable to what I call “the trickster problem.” To see the problem, suppose Madoff has found that emphasizing that a deal will help him in some way makes his potential victims more likely to take the bait. He tells a particular potential victim that he has a great investment opportunity and most of the capital already raised to go forward, but not quite enough. Madoff makes it clear to his mark that he needs her money to make the deal happen, or their great opportunity will be missed. He beseeches her, on behalf of himself and his other investors, to invest in the scheme with him. The trickster problem is this: Madoff might be interpreted as believing that because his potential victim will be directly and favorably moved by the thought that he is counting on her, that her good will and competence will extend to cover the domain of their interactions. Since the above example is based on the way he actually operated, it is clear that Madoff could have these attitudes toward his potential victim. Therefore, our working account of trust would have it that Madoff is trusting his potential mark to go along with his scheme. This is surprising and problematic because this case then runs
afoul of one of the canonical assumptions about trust: that trust that is violated tends to produce, and makes fitting, feelings of betrayal.

Though the case described above fits our current account of trust, it does not seem correct to say that Madoff would feel betrayed if his potential victim turned him down. He would most probably feel disappointed, but regardless of what he might feel, it strains credulity to say that it would be appropriate for Madoff to feel betrayed. We expect him to simply shrug his shoulders and move on to find his next victim. If he does not, but instead rails against his betrayal at the hands of his potential victim, we would think he was being unreasonable.

Our current proposal holds that when we trust another person, we believe that because she will be directly and favorably moved toward us, we can count on her good will and competence governing our interactions in a particular domain. But that proposal does not, apparently, include everything we mean to capture when we speak of trusting another person. In our example, Madoff expects that the good will and competence of his potential victim will be operative in their interactions because he believes she will be directly and favorably moved by the fact that he is counting on her, but we do not believe that he trusts her. It seems he merely relies on her to do what he wants, even though he satisfies the conditions of our current account of trust. The trickster problem, as exemplified by Madoff, shows that our current account of trust is mistaken.

3.6 The Participant Stance

The most prominent attempt to solve the trickster problem involves appealing to an idea developed by P.F. Strawson in “Freedom and Resentment.” Recall that part of
Strawson’s argumentative strategy in that piece involves drawing our attention to what he calls the *reactive attitudes*, a class of attitudes that we typically have toward persons but not toward nonpersons. As we saw in the previous chapter, Strawson’s reactive attitudes are emotions like “gratitude, resentment, forgiveness, love, and hurt feelings,” but the current proposal extends Strawson’s idea to other psychological states. Trust, on this account, is also a reactive attitude.

Pamela Hieronymi has insightfully explored Richard Holton’s idea that trusting someone involves a particular “stance” toward that person. It is generally agreed, as Hieronymi explicitly states, that the participant stance is the stance we take toward other responsible agents—that is, persons—and that this stance is inappropriate toward nonpersons. But reliance, we have noted, is simply characterized by the expectation that something will happen. Certainly it can be appropriate to rely on nonpersons or even inanimate objects to do certain things. Also, it is clear that we can simply rely on, rather than trust, other persons. What would the participant stance add to reliance to create trust, such that the resultant attitude would be inappropriate with respect to nonpersons?

What would it mean to expect a person to do something *from the participant stance*?

---

81 I focus here on Hieronymi’s discussion rather than Holton’s because Holton’s account doesn’t go very far toward explaining the phenomenon in question. Holton thinks of trusting someone as being vulnerable to betrayal in your interactions with them. On this account, when someone feels betrayed, that person’s betrayal is a reactive attitude. Disappointment, it seems, is not a reactive attitude for Holton.

Holton notes that while this distinction doesn’t provide a reductive analysis of trust, it allows us to point to a distinction between cases where there is trust and cases where there is no trust. But this is little progress, regardless of whether there is a reduction of the concept. For we have no explanation of why feeling betrayed is appropriate when one’s trust is violated or why betrayal, but not disappointment, is a reactive attitude.
Hieronymi’s tentative answer to these questions regarding relying on someone from the participant stance involves appealing to an analogy between two different grounds you might have for a belief that someone will do something. Consider, on the one hand, one kind of evidence you might have that someone else would do something: that person’s stated intention to do that thing. If someone asks why you believe that your friend Jim will mow his lawn, it makes perfect sense to say that Jim told you that he intends to mow it.

Now, it is possible to regard the intentions of other people as evidence in favor of certain predictions about what they will do. And clearly, as the case of Jim’s mowing illustrates, we sometimes do this. But, perhaps surprisingly, it does not make sense for us to do this in our own case. Following Richard Moran, Hieronymi notes that our own intentions to do things do not serve as evidence for us that we will do what we intend. When we believe that we will act as we intend to act, our belief is not based upon our intention, itself, for our intentions are not stable facts about us from our own deliberative perspective. Since it is up to us what we intend, Hieronymi argues that we should not regard our intention as evidence of how we will act, as it is always possible that we will change our minds about what to do. Suppose you are idly wondering whether you are going to mow the lawn today. You will not be able to answer this question for yourself by simply noting that earlier you formed the intention to mow. Now that you are attending to other countervailing considerations like how hot it is outside (and any other reasons for and against mowing), your estimation of the reasons for and against mowing should come to the fore while your previous intention should recede into the background.

Hieronymi exploits this difference in plausible available responses to the question “why should I believe that an agent will perform a particular action?” in her account of the difference between trust and reliance. When you trust someone, she believes, this involves thinking that another person will do something because of the reasons you believe her to have “in something like the way that one might form a belief about one’s own future on the basis of one’s own practical reasons.” Reliance, in contrast, is distinguished from trust by lacking this distinctive focus on the other person’s reasons.

3.7 Problems with the Participant Proposal

Suppose, then, that a unique focus on another person’s reasons defines the participant stance. But it is not clear how to vindicate Hieronymi’s idea that mere reliance on another person does not involve the participant stance, while trust does. This is because many cases of simply relying on another person will involve an implicit reference to that person’s reasons, given that we are dealing with another agent. For example, I rely on other motorists to allow me to proceed first when I arrive at a 4-way stop before them. If I consider why I rely on them, it is because I have certain beliefs about their reasons: I believe they want to avoid accidents and being cited by the police, and that these reasons are sufficient to get them to follow most traffic laws. But simply bringing the content of the reasons of others to mind does not amount to trusting them. Nor would it be reasonable for me to feel betrayed if one of the other drivers broke the law and entered the intersection before me. It might be reasonable for me to be angry

83 Ibid., 226.
with the driver because of what she did, but not to feel betrayed.

Perhaps the idea is, instead, that trusting involves a special kind of justification for my beliefs about the other person’s reasons. This way of developing the idea is supported by Hieronymi’s appeal to an analogy with testimony. When you testify to me that something is true, you offer a kind of guarantee. You put yourself on the hook to some degree, morally speaking, for what results from my new belief based on your testimony. Further, you personally stand behind my belief in a way that you would not if I came to believe by some other manner. Hieronymi might then object, based on the appeal to testimony, that the personal assurance that is characteristic of trust is not present in my driving example. I have beliefs about what the other driver’s reasons are, to be sure, but they are not based on any specific information about their beliefs. Perhaps if my belief about the content of their reasons was based on their testimony, or some other kind of uniquely personal assurance, then I could be said to trust the other drivers.

But this proposal also does not seem sufficient for trust. Imagine that I am hired by the city for a compliance project: I am to stand at 4-way stops and question drivers who stop about why they do so. Most drivers tell me that they stop because of a fear of getting a ticket and because they hope to avoid accidents. On the way home after a day of interviews, my path takes me to a 4-way stop. To my surprise, I notice that someone who answered my questions earlier in the day is stopped there as well. Even in recalling the fact that this subject told me that she obeys traffic laws because she is concerned

---

84 While I am skeptical that the appeal to testimony can directly address the trickster problem, the idea of “being on the hook” for something or “standing behind” something or someone has a normative ring to it that has an affinity with my preferred account. See sections 3 and 4.
about accidents, I do not trust her. If she starts into the intersection even though I have
the right of way, I certainly might become angry with her, as I might with any driver who
cuts me off, but I would not feel betrayed.

Further difficulty for understanding trust as reliance from the participant stance
comes from attempting to use it to solve the trickster problem. When Madoff relies on
his potential victim to go along with his plan, we can suppose that he has many true
beliefs about the kinds of considerations that she tends to view as reasons. As a good-
hearted person, she tends to support the plans of others people and she has even more of a
soft spot for friends of her friends. Madoff knows that she tends to be frugal with her
money, but also will take a chance from time to time on a more risky investment.
Further, he correctly surmises that she finds his current list of clients very impressive,
which gives her confidence in his excellence as an investment advisor.

All of these considerations make it quite reasonable for Madoff to think that his
potential victim will go along with his scheme so long as she doesn’t know that the
scheme is fraudulent. But even if Madoff is relying on her because he is making
reasonable assumptions about her reasons, and even if his victim has told him that these
are her reasons for participating, that simply indicates to him that she may be an easy
mark for his scam. None of these facts would warrant Madoff feeling betrayed if she
refused to invest with him.

These observations suggest that relying on another person because of personal
knowledge of the content of her reasons is not sufficient to create trust, nor is it sufficient
to warrant betrayal should the reliance be let down. This is true even when the other
person has testified to you about her reasons for doing what you are relying on her to do.
Feeling betrayed would not make sense toward the driver at the 4-way stop, nor would it make sense for Madoff to feel betrayed when his potential victim doesn’t go along with his plan. On Hieronymi’s account, it would seem that I trust the other driver and Madoff trusts his victim, therefore supporting feelings of betrayal if Madoff and I are let down. But as we’ve seen, this is at odds with plausible intuitions about the cases. Because that account is also not sufficient to solve the trickster problem, we are forced to conclude that there is a problem with Hieronymi’s account of trust as reliance from the participant standpoint.85

3.8 Feeling Betrayed

A clue to developing an account of trust that will solve the trickster problem can be gained by reflecting on what it is to feel betrayed. Since Baier’s seminal paper, theorists interested in trust have been appealing to the idea that when someone’s trust is violated, it makes sense for that person to feel betrayed, not to merely feel disappointed. But theorists working on trust have not asked what the difference is between feeling betrayed and feeling disappointed. Here is a proposal that makes sense of the distinction and also has important ramifications for an account of trust.

Both feeling betrayed and feeling disappointed involve emoting in a particular, negatively tinged way. As understood in the current context, both feeling betrayed and feeling disappointed are emotional phenomena—they necessarily involve felt instances of emotion, not merely dispositions or traits to feel something or respond in some manner.

85 It’s also possible that connecting trust to the participant standpoint is correct, but that Hieronymi’s account of that standpoint is problematic. I want to set aside this possibility to focus on what I think is a more promising direction of analysis.
But betrayal and disappointment are not basic emotional responses like anger, guilt, or joy. When a person “feels betrayed” or “feels disappointed,” her underlying emotional state is consistent with a plurality of emotional responses, including anger, disgust and sadness. It is perfectly coherent to feel betrayed or disappointed in virtue of feeling angry about what someone has done. It is also possible to feel betrayed or disappointed in virtue of feeling sad or disgusted about what has transpired.

These observations suggest that the distinction between feeling betrayed and feeling disappointed does not stem from the affective emotional experience that undergirds them. We would recognize a difference between disappointment and betrayal even in cases where the person who is disappointed and the person who feels betrayed both feel very sad because of what happened.

What distinguishes feelings of betrayal from feelings of disappointment is that the clearest cases of feeling betrayed involve feeling anger, disgust or sadness at a person’s action that ruptures an important relationship you have with her. Being disappointed lacks this specific content. We would not say that Elizabeth Edwards feels disappointed

---

86 The basic emotions are thought to make up a class from which all other emotions are derived. There is some dispute about exactly which emotions are basic, but a representative list might include some or all of the following: amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilt, pride in achievement, relief, sadness, satisfaction, sensory pleasure and shame. This list of 15 emotions if from Ekman, “Basic emotions,” 45-60. Ekman defended a shorter list of six basic emotions in earlier work (happiness, surprise, anger, sadness, fear, disgust/contempt).

87 In the course of feeling betrayed, a person might feel all of these underlying emotions concurrently, or serially, or only feel one or two of them. When Elizabeth Edwards learned of her husband’s affair, she apparently felt all three of these emotions in the course of feeling betrayed. “I cried and screamed, I went to the bathroom and threw up,” she writes. “Elizabeth Edwards Details Husband’s Affair in New Book - The Caucus Blog - NYTimes.com.”
that John committed adultery. We describe her as feeling betrayed because the harm she has suffered is specific to her relationship with John. We also describe Bernie Madoff’s clients as feeling betrayed because they took him to be more than just an investment advisor, but also their friend. Other things being equal, feelings of betrayal tend to be more intense, given that they are responsive to the breach of relationships that are centrally important to our lives. But the difference between feeling betrayed and feeling disappointed continues to assert itself even if we suppose that the intensity of the sadness is equivalent across an episode of betrayal and an episode of disappointment. This is because feeling betrayed involves an implicit reference to the external situation—the breach of a relationship due to the actions of one of the parties—that warrants the emotion.

3.9 **Relationships of Trust**

That feeling betrayed involves being emotionally bothered by the fact that an important relationship has been violated or ruptured by the actions of one of the parties points our account of trust in an important direction. It is no accident that the attitude that is characteristic of trust—believing that because a person will be directly and favorably moved toward us, we can count on her good will and competence governing our interactions in a particular domain—is an attitude that is a constitutive part of many different socially salient relationships.\(^88\) We expect that attitude from friends, romantic

---

\(^88\) The idea that the relevant relationship be one that is recognized as “socially salient” is meant to exclude relationships that hold between two people simply in virtue of sharing some property. For example, two people might share the relationship of having the same
partners, parents and others whom we count on. What has been missing from previous accounts of trust is that in the context of a relationship characterized by trust, we believe not just that we can expect good will from someone, but that we are entitled to it. It is our entitlement to that person’s good will, and the fact that we take the other party to owe it to us, that legitimates our feeling betrayed when trust is violated.

The language of “entitlement” and “owing” is natural here, but I want to contrast its current use with another use of “entitlement,” that is, a rule-governed institutional right. For a person to be entitled, in my sense, is for that person to have a normative claim on another person’s good will based on her participation in an interpersonal relationship, not an institution. The claim on someone’s good will based on the relationship is supported by norms that need not be explicit rules. However, the structure of the claim does bear some similarities with claims of institutional entitlement.

Let me say a bit more about how I understand the relevant interpersonal relationships. In a healthy friendship or romantic partnership, for example, the parties involved have standing dispositions to treat each other well, spend time with each other, and help each other. A healthy friendship between two people will also involve distinctive ways of caring and responding to each other. A true friend gets upset at the things that negatively impact you and cares about your projects. When you are a party to a relationship of this kind, one of your overarching presumptions about the other person is that she will act toward you with good will because she will be directly and favorably moved by the fact that you are counting on her.

---

However, it is not simply a happy detail that the other person bears you good will and that you realize this. In the context of a healthy interpersonal relationship, that the other person will act toward you with good will because she will be directly and favorably moved by the fact that you are counting on her becomes normative.\textsuperscript{89} That is why, if you are treated poorly by a friend, you can respond to the slight by saying “Aren’t we friends?” By doing this, you call your friend’s attention to the relationship the two of you share and how it gives you each reasons to foster and care for one another. By slighting you, your friend breaches the terms of the relationship. In order to make sense of this idea, we must take relationships like friendship to be partly normatively defined. It is not just that the parties to the relationship have attitudes that coalesce happily; they can come to have a normative claim on these attitudes from each other.\textsuperscript{90}

What is striking about the normatively constituted relationships I have been discussing is that one of the overarching norms of all of these relationships is what I call the trust norm. According to the trust norm, when you are a party to a normatively defined relationship of the right kind, the other person who is a party to that relationship owes you good will because she should be directly and favorably moved by the fact that

\textsuperscript{89} I take no stand here on what explains the fact that we are sometimes entitled to another person being moved in the right way to act toward us with good will. The account of trust I offer is neutral between an account of interpersonal relationships that is “voluntaristic” in treating the relevant entitlements as arising only from the voluntary actions of people and an account of these relationships that allows for legitimate claims on the good will of others being supported by considerations other than voluntary actions. For more on the distinction see Ibid., 191.

\textsuperscript{90} I draw the example of appealing to the terms of a relationship from Margaret Gilbert, A Theory of Political Obligation: Membership, Commitment, and the Bonds of Society (New York: Oxford University Press, 2006), 149-153. Gilbert’s target analysis is the notion of a joint commitment, rather than trust, though she notes there are important similarities. I follow her in thinking that trust, like a joint commitment, is normative.
you are counting on her. In order to qualify as a proper friend, romantic partner or parent, you must see the other party as having a claim on your good will. In this way, these relationships presuppose trust.

3.10 The Trickster Problem, Solved

This account of normatively defined relationships can help us solve the trickster problem. It explains why Madoff does not trust his potential victim to do as he wishes and why it will not be fitting for him to feel betrayed if his mark does not go along with his scheme. In the context of his scheme Madoff is not entitled to his victim’s good will toward him. The flip side of this claim is that Madoff’s victim does not owe him her good will.

Key to solving the trickster problem and making sense of our intuitions about the previous examples is that trust involves the belief that the good will of another person will cover our interactions with her because she owes us good will due to the terms of our interpersonal relationship with her. Madoff is not actually coming to his victim under the umbrella of an interpersonal relationship that involves trust. He may well be presenting himself that way — and, indeed, his presentation of himself as a trusted friend was a large part of the explanation of his success. Madoff’s pretense at friendship attempts to lead his potential victim to act because of her good will toward him, not merely to act because of a potential benefit to herself. Given that he comes to his potential victim as a friend, it may appear to her that she has some reason to be motivated

---

91 “Terms” can mislead. I do not assume that the relevant terms need to be explicitly stated or agreed to.
by the fact that he is counting on her to act with good will toward his plans. Since we can imagine her being persuaded of his good intentions toward her under the guise of friendship, we can imagine that Madoff can then have the beliefs characteristic of trust that involve relying on her good will and concern for him. But while he can have reasonable beliefs about his victim’s attitudes, he is not entitled to treat her attitudes with the significance they would normally have for a friend. He is not entitled to believe something further—that his victim owes him her good will. And since he realizes he is not entitled to treat her attitudes as characteristic of friendship, he should not feel betrayed if she does not go along with his plan.

3.11 Preliminary Conclusion

In this chapter, I have offered an account of trust that aims to solve the trickster problem. It does this by calling attention to a neglected feature of trusting. Some previous accounts of trust tried to give a specification of the requisite mental attitudes on the part of the truster that, taken together, would add up to trust. Other accounts appealed to the notion of the participant stance. Neither, on its own, could solve the problem.

My account is, in some ways, a hybrid of these two approaches. I also appeal to mental attitudes on the part of the truster, namely a belief that the other person will act toward you with good will because she is directly and favorable moved by the thought that you are counting on her. But that is not enough, as the trickster case exemplified by Bernie Madoff demonstrates. The missing ingredient is normative. Someone who reasonably trusts another has a legitimate normative claim on the good will and concern of the trusted person.
To explain this claim, I have appealed to a number of interpersonal relationships, including friendships and romantic partnerships. It is important to emphasize that I do not exclude other interpersonal relationships from providing the normative backing that is required for trust. All that is required is that a requisite relationship includes, as one of its constitutive norms, what I call the Trust Norm. The trust norm states that

Because it is partly constitutive of a normatively defined relationship you are a party to, the other person who is a party to that relationship owes you good will because she should be directly and favorably moved by the fact that you are counting on her.

If a relationship includes the trust norm as one of its constitutive norms and two people are parties to that relationship, then the normative backing of that relationship will entitle at least one of the parties to the good will and concern of the other. In many such relationships, of course, the relevant trust norm is symmetric.

The trust norm does the relevant work in solving the trickster case since it shows that a trickster should not count on the good will of her victim as the trickster’s victim does not owe him good will and concern. But it also helps to explain the phenomenon of feeling betrayed, as contrasted with feeling disappointed. Feeling betrayed involves feeling violated, feeling that the relationship that you were entitled to count on was breached by the actions of the other party. This phenomenon makes sense in the context of a relationship that is normatively defined and includes the trust norm as one of its constitutive norms.

But the trust norm is not all there is to trusting. You could be entitled to count on the good will of another person because of your participation in an interpersonal relationship with her, but not believe that she will come through. This is just the situation
that Elizabeth Edwards faces in response to John’s infidelity. She knows that she is entitled to John being directly and favorably moved to act toward her with good will, but she is not at all sure that he will act in that way. To the extent that Elizabeth Edwards doubts that John will act because of good will toward her, she fails to trust him.

This is why the correct account of trust requires both normative and nonnormative attitudes. We need the normative part to exclude Bernie Madoff from counting as trusting his potential victim and we need the nonnormative part to properly characterize Elizabeth Edwards’ struggle to trust John. Our final account of trust holds that

To trust someone is (i) to believe that because a competent person will be directly and favorably moved toward us we can count on her good will governing our interactions in a particular domain and (ii) to believe that we are entitled to her good will because we are a party to a normatively characterized relationship with that person.

Trust, then, is a rather complex attitude. It is no wonder that it has taken so long to get it right.

3.12 The Relevance of Trust to Moral Responsibility

While I’ll return to this issue in more detail at the conclusion of the next chapter, I want to take a moment here to situate this chapter in the complete dissertation. Recall that the overall motivation for the project is a consideration of whether people like Professor Plum who have been globally manipulated are morally responsible for the actions they take. I believe that they are, but showing this requires attention to both emotional and nonemotional reactive attitudes. Trust is a central case of the latter type of attitude, but its connection with attributions of moral responsibility has not always been clear. My analysis of trust provides that clarity, because if trusting depends on beliefs
about a person’s good will and whether we are entitled to her good will, we can see how the degree to which it is fitting to trust someone will often depend on how she acts—the degree to which it is fitting to trust her will depend on whether or not her actions display good will. In the next chapter, I’ll say more about how trust fits into the overall project and about whether it is fitting to trust Plum, given that he killed Ms. White.
4.1 Introduction

In the previous chapter, we met an account of trust adequate to handling the trickster problem. That account suggests that trust is a moral responsibility-tracking attitude just like anger, but we have yet to see trust situated within a more general account of holding someone morally responsible that makes its connection to moral responsibility clear. This chapter takes that as its main aim.

This is important because, while trust is a central attitude for understanding people’s moral responsibility, it is not the only nonemotional attitude that is implicated in fully understanding what it is to hold someone morally responsible. This can be illustrated with a comparison to the reactive emotion account of moral responsibility we examined in Chapter 2, where our reactive emotion account held that if anger is fitting for someone, then she is morally responsible. But anger is not the only reactive emotion. Gratitude is reactive and other emotions are likely reactive as well (determining this requires close attention to the candidate emotion’s evaluative presentation). I was content to primarily focus on anger because my aim in this project is to show that Professor Plum is morally responsible for killing Ms. White. Showing that anger is fitting for Plum is enough to satisfy that aim, but a full version of a reactive emotion view of moral responsibility would delineate all reactive emotions. Similarly, a full account of nonemotional reactive attitudes must offer a characterization of attitudes other than trust.
In the remainder of this chapter, I lay out such an account and show that, on it, Plum is morally responsible for killing White.

I begin with a return to Strawson’s “Freedom and Resentment.” I show that when contemporary Strawsonians develop an account of moral responsibility in terms of the reactive attitudes, they actually diverge from Strawson’s own understanding of moral responsibility. This helps to motivate the idea that a view that understands moral responsibility by appeal to attitudes other than emotions has a legitimate claim on a Strawsonian pedigree.

T.M. Scanlon’s account of holding someone morally responsible in terms of nonemotional attitudes is the most developed such account, so I take it up in the remainder of the chapter. Scanlon develops his view via an analogy to nonemotional attitudes that are characteristic of, and help constitute, our interpersonal relationships with other people. For example, friends believe that they can count on each other and they also intend to spend time together and help each other out. Scanlon argues that when a friend does something that runs afoul of what we take ourselves to properly expect, then it may make sense to modify the extent to which we intend to spend time with that person or confide in her. On Scanlon’s theory, when we make that modification in our attitudes toward our friend because how she acted shows us something important about her attitudes toward us, we hold her morally responsible.

Scanlon also thinks of morality as a kind of relationship that holds between persons.

92 “Freedom and Resentment.”
However, a problem lurks in his attempt to extend his relationship-based account to cases where we have no previous attitudes toward a person. This is because some of our moral attitudes are nonoptional and so cannot be permissibly modified by the actions of others. Other moral attitudes are optional, but we are unlikely to have them toward people with whom we are not yet in a relationship. This means that with respect to some people, like a stranger, it is unclear whether Scanlon’s account succeeds in describing any attitudes that may permissibly be modified by the stranger’s blameworthy action.

I want to resolve this bit of opacity in Scanlon’s account by emphasizing the centrality of trust for understanding what it is to hold someone morally responsible. This is because, while trust and distrust are contrary attitudes, they are not contradictory. So it is not required that you already trust another person in order for distrust of that person to be a modification of your attitudes toward her.

I then show that this is the way we should think about the significance of Plum’s action when he kills White. Regardless of whether we trusted him beforehand, his action gives us good reason to distrust him. As trust is a centrally important attitude for the reactive nonemotion account of moral responsibility we should conclude that on such an account, Plum is morally responsible.

4.2 NONEMOTIONAL REACTIVE ATTITUDES

Human beings have a range of mental attitudes: beliefs, desires, emotions, wishes, etc., but only some subset of these are reactive attitudes. Many philosophers take the Strawsonian view of moral responsibility to hold that a person is morally responsible when she is the appropriate target of reactive emotions, exemplified by praise and blame.
But Strawson’s followers have been too quick to identify Strawson’s own conception of moral responsibility with the appropriateness of emotions, to the exclusion of all other attitudes.

To see this, let us return to “Freedom and Resentment.” Strawson does not begin his attack on pessimism by plunging directly into a discussion of the compatibility of determinism and moral responsibility. Instead, Strawson says that he wants to set aside the issue about the compatibility of moral responsibility and determinism and turn to an examination of something else:

I have mentioned punishing and moral condemnation and approval; and it is in connection with these practices or attitudes that the issue between optimists and pessimists—or, if one is a pessimist, the issue between determinists and libertarians—is felt to be particularly important. But it is not of these practices and attitudes that I propose, at first, to speak. [...] I want to speak, at least at first, of something else: of the non-detached attitudes and reactions of people directly involved in transactions with each other; of the attitudes and reactions of offended parties and beneficiaries; of such things as gratitude, resentment, forgiveness, love, and hurt feelings.94

Strawson’s methodology is to set aside the issue that concerns the optimists and the pessimists—whether moral responsibility is possible in a deterministic world—and to focus instead on what he takes to be a related issue: whether the reactive attitudes are appropriate (as I would say, fitting) for us in a deterministic world.

Yet despite the fact that so many philosophers who count themselves as Strawsonians understand moral responsibility claims in terms of the appropriateness of reactive emotions, Strawson does not strictly equate holding someone morally

---

94 “Freedom and Resentment,” 62. Interestingly, Strawson rarely uses the language of “moral responsibility.” He more commonly speaks of “moral condemnation and approval,” “moral indignation” and “moral disapprobation,” but he makes no attempt to carefully distinguish these ideas conceptually.
responsible with feeling reactive emotions. When Strawson speaks of moral
responsibility, he is not speaking of the reactive attitudes but of a “sympathetic or
vicarious or impersonal or disinterested or generalized analogue […] of the reactive
attitudes.” Strawson clarifies his remarks by saying that the attitudes he wants to
delimit are not essentially vicarious, but are essentially capable of being vicarious. His
point seems to be that the nonvicarious reactive attitudes are felt or experienced by a
person in situations where things of relatively direct or immediate concern to that person
are impacted. Feeling anger or resentment is perfectly normal if another person harms
my partner, my cat, or my car, because I care about all of these things. But suppose I
read about the way a politician embezzled money in a political jurisdiction far from my
own. If I feel no particular connection to this politician or his political wrong, I might not
feel anger at the perpetrator or even think about whether anger is appropriate for him.
Instead, I might simply disapprove of what he did.

So, Strawson seems to identify moral responsibility with a class of “vicarious” or
“impersonal” attitudes. What could these attitudes be? It is not totally clear from
Strawson’s discussion. Strawson says that the vicarious moral analogue of resentment is
moral indignation or disapproval. What makes it moral, he claims, is that your own
interests are not involved. But he then goes on to claim that you can feel this attitude on
your own behalf, and it is unclear whether or not the attitude in question is an emotion.
Therefore, Strawson’s discussion leaves it open that someone might explicate a
conception of moral responsibility that takes nonemotional attitudes to be reactive
attitudes. Such a position will bear more similarity to Strawson’s official position in

95 Ibid., 70.
“Freedom and Resentment” than is typically allowed. If this is right, then the account of moral responsibility offered in this chapter has a claim on being a “reactive attitude” account of moral responsibility. Scanlon has developed such a view, so to his account we now turn.

4.3 Scanlon’s Theory—Personal Relationships

Scanlon has developed a theory of holding people morally responsible that emphasizes nonemotional attitudes like expectations and intentions. Scanlon accepts that reactive emotions have a place in our moral responsibility practices, but holds that an account of holding people morally responsible only in terms of the reactive emotions would be “too thin” because it would fail to take account of other modifications in our attitudes that occur when we hold someone morally responsible. This is because Scanlon wants to understand our moral responsibility practices by situating those practices within an account of interpersonal relationships. While being disposed to have certain emotions is an important aspect of our relationships with other people, Scanlon takes intentions and expectations to be the central attitudes for understanding these relationships.

The case of friendship presents an instructive analogy with the moral relations

---

96 Pamela Hieronymi defends a similar account of the significance of blame in her article “The Force and Fairness of Blame,” Philosophical Perspectives 18, no. 1 (2004): 115-148. For ease of exposition, I focus primarily on Scanlon’s account.

97 Moral Dimensions: Permissibility, Meaning, Blame, 143.

98 Ibid., 131-2. Scanlon speaks of blame, rather than holding someone morally responsible, but his discussion makes it clear he is concerned with the same topic. An earlier, and less complete, specification of his views is in Chapter 6 of What We Owe To Each Other.
Scanlon has in mind:

To be friends with a person involves such things as intending to give help and support when needed, beyond what one would be obligated to do for just anyone; intending to confide in the person and to keep his or her confidences in return; and intending to spend time with the person when one can, and to “keep in touch.” Being a friend involves actually being disposed to act in these ways, not just having an abstract intention to do so. It also involves being disposed to do these things for the right kinds of reasons, not only out of a sense of obligation but also out of a certain kind of concern and affection for each other.99

The flip side of these claims involves corresponding expectations. If you are a party to a mutual friendship, you expect your friend’s support when needed; you expect her to confide in you and to keep your confidences, etc. And you expect her to do these things for reasons based on her concern and affection for you.

Suppose, now, that you learn that a friend of yours—Joe—was at a party making cruel jokes at your expense that involved some revealing facts about you that you had told him in confidence.100 Joe’s action urges you to consider whether you and Joe are still friends, and, if you are, what kind of significance his actions have for your friendship. Scanlon identifies coming to a conclusion about the significance of Joe’s actions for your friendship with judging Joe morally responsible for his actions.101 But there are other kinds of reactions that you might have toward Joe that Scanlon believes constitute holding Joe morally responsible. You might revise your attitude toward Joe in light of your conclusions about your friendship (deciding to spend less time with him, for example). Finally, you might go a step further and express this change in attitude to Joe.

100 The example is Scanlon’s. Ibid., 129.
101 Scanlon speaks of Joe being blameworthy, rather than Joe being morally responsible, but it is clear he means the same thing.
Using the terminology set out in Chapter 1, we can say that coming to the conclusion that Joe’s actions are significant for your friendship amounts to thinking that Joe is morally responsible for what he did. Going further than that in actually revising your attitudes toward him or expressing these attitudes involves holding Joe morally responsible. Joe is morally responsible when he has performed an action that has significance for the relationships you can have with him. Finally, expressing a change in attitude toward Joe would be to hold him accountable.

4.4 Scanlon’s Theory—Moral Relationships

Scanlon wants to make similar claims about the way in which our moral relationships can be impacted by the actions of other people. Let’s use the term “relationships of mutual recognition” to refer to the moral relationships that Scanlon has in mind. Scanlon thinks that morality requires us to take care to not harm others, to help them when we can do so easily, not to lie to them, etc. and that these requirements constitute a kind of relationship. When two people recognize that their status as rational

---

102 One interesting result of Scanlon’s view is that it treats us as thinking people are morally responsible and holding them morally responsible when some might not think this plausible. For example, suppose I have an alcoholic friend who is unable to keep commitments to me because of his alcoholism. On Scanlon’s account, if I decide that because he cannot keep commitments, I am unable to continue spending time with him, I both think and hold my friend responsible.

Note that thinking and holding my friend responsible for his inability to keep commitments is consistent with also seeing him as a victim because of his alcoholism. In such a case we are led to two assessments of the friend that are in tension with one another, just as saw there was a tension between the assessments of Plum in Ch. 2 as someone for whom anger is both fitting yet undeserved. I believe we should draw the same conclusion about the alcoholic friend who cannot keep commitments as we did for Professor Plum: this is someone who is morally responsible for what he does, and we should have sympathy for this person.
beings requires this kind of moral concern, they stand in the relationship of mutual recognition. If a person acts toward you in a way that is inconsistent with recognizing your moral status, that has implications for whether you should take yourself to have a relationship of mutual recognition with that person.\textsuperscript{103}

For example, suppose that a stranger stops you in the street and, brandishing a weapon, demands that you give him all of your money. (Assume that the stranger doesn’t show appropriate concern for your rights and that his action was not justified by any of his circumstances.) Fearful, you comply, handing over money before you run off. After you are home, you begin to reflect on your relationship with the stranger. Scanlon would suggest that when you initially met the stranger, you thought that your relationship with him might be characterized by mutual recognition. However, his demands and subsequent actions dash you of that hope. The stranger’s actions indicate to you that he does not regard your moral standing as providing him with sufficient reason to forbear from certain courses of action. Therefore, Scanlon would urge, it makes little sense for you to regard him as someone with whom you can sensibly try to foster mutual recognition. In such a situation, Scanlon would suggest that some of the morally significant interactions we have with people are inappropriate for the stranger. For example, you should come to the conclusion that you do not have any reason to help the

\textsuperscript{103} Compare Hieronymi: “I don’t resent you because you failed to achieve some standard of human excellence or failed to live up to the norms internal to willing. Rather, I resent you because your action or attitude communicated disrespect or disregard for me. Likewise I am not grateful because you performed some particularly fine specimen of human action; I am grateful because you chose to benefit me in a way that you needn’t have. I am more concerned with myself, or with the relation in which we stand, or with your concern for me, or with how I figure into your world, than I am concerned with your performance, as such—with how you fared against a standard of moral excellence or rectitude or goodness.” “The Force and Fairness of Blame,” 124.
stranger complete his projects should you have an opportunity to do so in the future.

4.5 PROBLEMS WITH SCANLON’S MODEL

I find Scanlon’s idea suggestive, but as he describes it, I have concerns about the ability of the account to capture how we hold people morally responsible in a significant number of important cases, including that of Professor Plum. As Scanlon notes, accounts that treat holding someone morally responsible in terms of the appropriateness of nonemotional attitudes are sometimes accused of offering an account of holding someone morally responsible that does not explain the distinctive weight that moral blame seems to have. Unless we say more about why we are interested in this kind of character assessment, it may seem to be a pointless assessment of moral “grades.”

While it is not always clear just what the distinctive “weight,” or significance, a grading account of holding people morally responsible is supposed to lack, it is often thought that an account of moral responsibility that focuses on the reactive emotions does not suffer from this defect. Scanlon traces the “weight” of an emotional account to the fact that part of being involved in normal interpersonal relationships is to be susceptible to the reactive emotions as a result of the good or ill will that others show toward us.

Scanlon’s own account of holding someone morally responsible, which emphasizes changes in intentions and expectations, also relies on our participation in normal interpersonal relationships to explain the significance of holding people responsible, so it appears to avoid the charge of being simply a matter of “pointless moral grading.” But

---

there is still a problem for Scanlon’s account, since an account in terms of the reactive emotions appears to better explain the phenomenon of holding strangers morally responsible.

To see this, return to the above example involving being robbed by a stranger. While in the moment, it would be common to simply feel fear at the prospect of being robbed, it is evident from Chapter 2’s discussion of anger that feeling anger at the stranger would be fitting, so the stranger is morally responsible. Further, since it is reasonable to think that your anger at the stranger is a response to the significance of what the stranger has done, and we typically find the reactive emotions that others have toward us to be significant, an account of holding someone morally responsible based on the reactive emotions can easily hold strangers morally responsible in a way that appears to have the requisite “weight.”

Consider now what Scanlon’s account could say about the same case. The significance of holding someone morally responsible, on his account, is to be found in a revision or modification of your attitudes toward someone.105 As such, holding your disloyal friend Joe morally responsible can be understood as finding that his actions require a revision of your already existing friendship. For example, you might decide it’s necessary to alter your standing intention to spend time with him and help foster his projects.

But what already extant attitude toward the stranger who robbed you could be similarly revised? As Scanlon notes, there is a set of intentions toward others that morality requires us not to have. For example, it is always wrong to intend to kill or

---

harm others unless certain special circumstances, such as self-defense, apply. As such, the exception for self-defense is built into the intention not to harm others, so we don’t modify this intention when we act in self-defense toward an aggressor. Therefore, these intentions may not permissibly be revised in light of the fact that the stranger has shown ill will. Even worse, the stranger’s action no longer presents a threat\textsuperscript{106}—the robbery is over and done with—so you no longer have a defensive justification that would permit you to intend to harm the stranger.

For us to understand the significance of holding someone responsible on Scanlon’s account, we need to be able to make sense of a way in which other aspects of our moral relationships can change as the result of a stranger’s ill will. Scanlon explains this by appealing to the fact that when someone has shown ill will, it can be appropriate to refuse to enter into agreements with that person or to rule out the possibility of friendship with him. But as Scanlon himself notes, there is no general moral obligation to befriend everyone or enter into agreements with them.\textsuperscript{107} Therefore, we may not have such attitudes toward strangers in the first place so that they could then be modified (and, it is permissible not to have them).

Scanlon argues that his account avoids being merely a form of moral grading since, when someone shows ill will toward us, the appropriate modifications of intentions and expectations toward that person are really modifications of already existing, morally weighty relationships. Modifying a relationship with someone requires that there already be an existing relationship to modify. But in what sense is there such a

\textsuperscript{106}For simplicity, assume also that the stranger’s action has not created any future vulnerabilities to harm which you might prevent by harming him now.

\textsuperscript{107}Moral Dimensions: Permissibility, Meaning, Blame, 143.
relationship to be modified with a stranger? An appeal to basic moral requirements might make sense of the idea that there is a relationship, but these requirements are not modifiable. On the other hand, interactions that are not owed unconditionally to people are generally not present toward strangers; when they aren’t present, we can’t sensibly speak of there being a relationship to modify. If so, Scanlon’s account of holding people morally responsible is in trouble; it may be reduced to moral grading in the case of strangers. Even worse, for my purposes, it appears unhelpful for thinking about Professor Plum’s moral responsibility, since it is not obvious that Plum had an antecedent relationship with anyone, much less with Ms. White.

4.6 A WAY OUT?

One option for Scanlon is to simply allow that his account of holding morally responsible in terms of the revision of intentions and expectations characteristic of interpersonal relationships fails to hold the stranger morally responsible. He could allow this because, while he wants to emphasize the modification of intentions and expectations in his account of holding morally responsible, he does believe that part of holding someone morally responsible involves the reactive emotions. If so, the fittingness of anger for the stranger could serve to explain the sense in which the stranger is held responsible, even though a nonemotional account has nothing to say about the case.

Scanlon emphasizes that holding some people, like close friends, morally responsible is generally more weighty than holding people responsible who are not already friends. This phenomenon might be thought to help Scanlon deal with the stranger. He is held responsible, but it is just not as important to us since we lack a relationship with him ahead of time. This won’t help Scanlon, though, since the current worry is that we might completely lack any relations that can be modified with the stranger before we meet.
This would be a mistake. Intuitively, something has changed in your relations with the stranger who robbed you other than the fact that it now is fitting to feel anger at him. Consider that if you saw the stranger again, you would likely shun him in a way that you would not with another stranger. And this might be true even if you were in a social context where it was clear that the stranger had no weapon and you were likely in no danger. We can make sense of this phenomenon, while at the same time helpfully filling out Scanlon’s account, by focusing on trust.

4.7 Trust to the Rescue

Scanlon’s account of the way in which we hold someone morally responsible foundered on the case of strangers because we could identify no previously present attitudes toward a stranger that could permissibly be modified. The attitudes that morality requires us to have toward all people could not be modified and the attitudes that can permissibly be modified were arguably not there.

Trust can help fill some of the gap in Scanlon’s analysis because trust has a contrary attitude—distrust—that is not also contradictory. That is, there is a neutral space between the attitudes of trust and distrust—it is possible to have neither attitude toward a person. As I explained in the previous chapter, trusting someone involves two components:

(i) Believing that because a competent person will be directly and favorably moved toward us, we can count on her good will governing our interactions in a particular domain.
(ii) Believing that we are entitled to that person’s good will in the given

---

domain.

While trust is understood as involving the belief that the other person will act toward us with good will, distrust involves a contrasting attitude: believing that the other person will act on reasons that are inconsistent with what morality requires (he will show ill will toward you) or that he will act on reasons that are morally neutral while ignoring morally relevant considerations he should have attended to (he will lack good will toward you). The neutral attitude between trust and distrust is arguably the attitude that we usually have toward strangers; call the neutral attitude an attitude of “cautious optimism.” Cautious optimism is the attitude we have when we hope that people will act toward us with good will but do not yet believe that they will. The fact that morality requires that strangers bear us some good will makes it fitting to hope that we can rely on their good will to govern our interactions with them, but we do not always believe that we can. When this is true, we neither trust nor distrust such people.

This gives us a ready way to understand how our relations with the stranger change when he robs you. Before the robbery, you held the attitude of cautious optimism toward the stranger. After the robbery, you realize that this is a person who is clearly not motivated by good will toward you, so the robbery makes it appropriate for your attitude toward the stranger to become one of distrust. This means that the stranger’s action has implications for your relationship with him. Even if it is incorrect to describe you as

---

10 As I’m considering cases that contain only competent adults, I omit discussion of the competency clause in what follows.

11 A careful analysis of hope would show us when hope is fitting and when it is not. I have no such analysis to offer here, other than to note that hope seems to have a lower standard for fittingness than belief: it’s not fitting to believe that my basketball team will win when they are down by 5 with 12 seconds to play, but it is fitting to hope that they will win because the possibility of winning still exists.
having had a prior attitudinal relationship with the stranger when you viewed him with
cautious optimism, when you come to distrust him, you move to having an attitudinal
relationship with him that you previously lacked. The relevant distrust relationship is
negatively tinged, rather than positive, as in the case of friendship. But it is a relationship
that matters, nonetheless.

Since whether it is fitting to trust or distrust someone is clearly a matter of
significant social importance, emphasizing the centrality of trust to a nonemotional
reactive account like Scanlon’s allows him to avoid the charge of offering an account of
holding someone morally responsible that is merely concerned with moral grading. Still,
it might be worth stressing a few ways the appropriateness of trust and distrust matters to
us. Friendship and many other important relationships require some level of trust. Many
forms of cooperative activity require that the cooperators trust each other to do what is
expected of them. Societies and governments where trust tends to flourish generally
function more efficiently. Even further, when people in a society are generally
trustworthy, that provides you with a reason to expect trustworthiness even from people
who have not independently demonstrated to you that they are trustworthy. The flip
side of this claim is that when people with whom you interact with are not trustworthy,
you have a reason to distrust others. The actions of a stranger therefore may have
implications not only for your trust or distrust of that person, but for your trust or distrust

112 Baier makes this point in “Trust and Antitrust,” 232.
113 Philip Pettit, “The Cunning of Trust,” Philosophy and Public Affairs 24, no. 3
114 This consideration and how it is supported is the topic of Philip Pettit’s “The Cunning
of Trust.”
of other portions of society. For all these reasons, the fact that a stranger robs you can be extremely significant for your relations with others, even setting aside the impact of any reactive emotions. Emphasizing how the actions of strangers can sensibly impact your trust in others, then, gives Scanlon’s account of holding strangers morally responsible a significance that goes beyond mere moral grading.

4.8 Distrusting Plum

Now that we have an understanding of how talking about trust can help Scanlon’s theory explain how we hold strangers morally responsible, it makes sense to return to the case of Professor Plum. We now need to see what Scanon’s account of holding someone morally responsible would say about Plum’s moral responsibility for killing Ms. White. Let’s again review Pereboom’s case:

Professor Plum was created by neuroscientists, who can manipulate him directly through the use of radio-like technology, but he is as much like an ordinary human being as is possible given this history. These neuroscientists manipulate him to undertake the process of reasoning by which his desires are brought about and modified. They do this by pushing a series of buttons just before he begins to reason about his situation, thereby causing his reasoning process to be rationally egoistic. Plum does not think and act contrary to character since his reasoning process is often manipulated to be rationally egoistic. His effective first-order desire to kill White conforms to his second-order desires. The process of deliberation from which his action results is reason-responsive; in particular, this type of process would have resulted in his refraining from killing White in some situations in which the egoistic reasons were otherwise. Still, he is not exclusively rationally egoistic, since he typically regulates his behavior by moral reasons when the egoistic reasons are relatively weak – weaker than they are in the current situation. He is also not constrained in the sense that he does not act because of an irresistible desire – the neuroscientists do not provide him

with a desire of this kind.\textsuperscript{116}

Before Plum unjustifiably kills White, it is fitting to have cautious optimism toward him. Plum is aware of, and responsive to, moral reasons, though he often fails to act on them when egoistic reasons are stronger. So it is fitting to hope that he will do as morality demands, even if it is not fitting to believe that he will do so. I have suggested that to distrust someone is to have an attitude toward him that is the contrary of trust: the clearest cases of distrust are those when we believe that someone’s ill will toward us will govern our interactions in a particular domain. Moving from cautious optimism toward someone to distrusting him, then, counts as holding someone morally responsible on Scanlon’s view. Do Plum’s actions, then, give us reason to move from an attitude of cautious optimism toward him to an attitude of distrust?

It is clear that they do. Plum’s action confirms what Pereboom has made explicit in the example: Plum generally weighs moral reasons, like the rights of persons, far below his own egoistic interests. While it isn’t clear from the example upon what egoistic reason Plum is acting, we can suppose that it is to get $200 in cash Plum knows White has. So Plum’s action tells us that he weighs the moral rights of persons far below his own interests in having what is a fairly small sum of money. Given what his action says about his deliberative perspective, which ignores moral reasons when they are in conflict with moderately strong egoistic reasons, we have good reason to distrust Plum.

Here is another way to see why distrust is fitting for Plum. Recall from Chapter 2 that emotions are often evaluated in a number of ways. Considerations of fittingness, moral considerations and prudential considerations might all be appealed to when

\textsuperscript{116} “Hard Incompatibilism,” 94-95.
adjudicating what to feel about what someone has done. A similar point can be made about trust and distrust. Trust is an attitude that consists of beliefs, not emotions, but different kinds of evaluations can be made of beliefs, just as they can be made of emotions. A objector defending Pereboom might argue that since Plum’s action was caused by the actions of the neuroscientists, it would be unfair to distrust him.

For the sake of argument, I am willing to grant this point to the objector. Suppose it is unfair for Plum to be distrusted by us. Still, that point does not impact whether distrust is fitting for Plum, in the sense elucidated in Chapter 2. There we said that an emotion is fitting just in case its evaluative presentation gets it right about the properties of the emotion’s object. The same point is relevant about trust. Trust is fitting, just in case the attitudes that constitute trust correctly describe their object. As a reminder, those attitudes are:

(i) Believing that because a competent person will be directly and favorably moved toward us we can count on her good will governing our interactions in a particular domain.
(ii) Believing that we are entitled to that person’s good will in the given domain.

Let us take (ii) first. Does our belief that we are entitled to Plum’s good will correctly describe our situation with Plum? It does. Plum is a rational agent and nothing Pereboom has said suggests that Plum does not act wrongly. Plum does wrong and Pereboom is perfectly willing to admit this.\textsuperscript{117} Given that he acts wrongly, we know that Pereboom takes normal moral demands (such as the rights of persons) to apply to Plum. General moral claims, then, give us an entitlement to a minimal level of good will from Plum, even though it is a level that he does not provide.

\textsuperscript{117} Ibid., 99.
So condition (ii) is satisfied. What about (i)? As we have already seen, Plum’s action suggests that, as a general rule, he weighs egoistic considerations more strongly than moral considerations. Killing White is an action representative of that attitude. This means that (i) does not characterize Plum correctly. What is supported is the contrary of trust: a belief that we can count on Plum’s ill will or absent good will governing our interactions. So distrust is fitting for Plum.

Compatibilists should again emphasize that, for the purposes of attributing moral responsibility to someone, fittingness conditions are the only relevant considerations of attitudinal appropriateness. So a compatibilist may grant that it is unwise, unfair, or just plain wrong to distrust Plum, but none of those claims is relevant to the moral responsibility-relevant claim, which is that distrust is fitting for Plum.

4.9 Initial Conclusion

In this chapter, I have argued that changes in our nonemotional attitudes like beliefs and intentions constitute an important aspect of holding each other morally responsible. I explored Scanlon’s account of this practice and found that while it does well at accounting for holding people responsible with whom we already have an interpersonal relationship, it has problems explaining just which nonemotional attitudes could be permissibly modified toward strangers as the result of their wrongdoing. I argued that supplementing the account with my discussion of trust from Chapter 3 helps the nonemotional account to avoid this problem by explaining that when a stranger wrongs

---

118 If (ii) was not satisfied then, in that sense, Plum would be out of the moral picture in a similar way to how we might regard a nonhuman animal that is incapable of understanding moral demands.
us, that makes it appropriate to regard him with distrust rather than cautious optimism. Finally, I argued that this is how we should view Professor Plum on a nonemotional account of moral responsibility. His killing of Ms. White makes it fitting for us to distrust him.

The emotional and nonemotional aspects of our practice of holding each other morally responsible have delivered a unanimous verdict: Professor Plum is morally responsible for killing Ms. White. So far, the manipulation argument has proven ineffective against a full account of what it is to hold someone morally responsible, and compatibilism is in the clear. But compatibilism is not fully vindicated yet, for we still need to consider what the results so far imply about the propriety of punishing Professor Plum. That is the project for the next chapter. But first, I want to briefly explain why both nonemotional and emotional attitudes are required to explain the ways in which we hold each other responsible.

4.10 A SYNTHESIS OF MORAL RESPONSIBILITY

It might be wondered, at this point, why I developed both the emotional and nonemotional strains of our practice of holding each other morally responsible when my overall task was to argue that Professor Plum is morally responsible for killing Ms White. Since discussing the fittingness of reactive emotions like anger allowed me to show that Plum was morally responsible, why bother talking about the nonemotional attitudes? Or, since my discussion of trust and its significance as a moral responsibility-tracking attitude also generated the same result, isn’t there a way in which the discussion of the reactive emotions was superfluous?
No; neither account would be complete on its own. An account of holding people morally responsible that left out either the emotional or nonemotional reactive attitudes would fail to fully characterize our practices of holding each other morally responsible and accountable. To show that both elements are necessary for a full account of Professor Plum’s moral responsibility, I’m going to be make two points. First, without an appreciation of the various attitudes involved in holding someone morally responsible, we would lack an understanding of why and when our behaviors toward someone count as holding her accountable. Second, an additional reason to focus on both emotional and nonemotional attitudes is that we often take the attitudes implicated in our moral responsibility practices to provide us with reasons to engage in holding each other accountable in various ways. If that’s true, then without a full account of the moral responsibility tracking attitudes, we’ll lack a full understanding of the reasons for, and justification of, our ways of holding each other accountable.

To begin to see this, focus for the moment on what we earlier called holding someone accountable, that is, our prototypic outward responses toward someone we hold morally responsible. Consider, for simplicity, only the reactions we have toward someone we regard as having done something wrong. We can divide those outward responses into four categories: anger expressions, hostile behaviors, reproaches and punishments. Anger expressions concern behaviors that communicate that someone is feeling angry toward you—making an angry face, for example. Hostile behaviors are responses that are normal behavioral upshots of anger that are not essentially expressions

---

119 Here I am influenced by George Sher’s analysis of blame. See In Praise of Blame (New York: Oxford University Press, 2006), 94-95.
of anger. Examples include throwing something at someone, attacking someone or even angrily writing someone out of out your will. Reproach involves intentional behavioral responses that are essentially meant to communicate our anger to the person in question. Examples include berating or pestering someone. Punishments involve the intentional administration of harsh treatment on the basis of past wrongdoing, but need not be administered with hostility.\footnote{This claim is not contradicted by Feinberg’s argument that punishment expresses anger or blame toward criminals. My point is that legal and nonlegal punishment might carry this emotional content even while it is administered dispassionately. “The Expressive Function of Punishment.”} For example, after your child drove recklessly, you might refuse to allow him to use your car without either experiencing or expressing anger toward him. Cases where we respond behaviorally to people with all four of these behaviors are exemplar cases of holding people morally responsible. Consider this example: you are fiddling with your MP3 player on the drive home from work. Not paying sufficient attention to what is in front of you, you have to slam on your brakes to avoid hitting a pedestrian in the crosswalk. The pedestrian turns to you with an angry glare, threateningly approaches your car and yells “What the hell are you doing?” After you apologize profusely, he calmly turns away, takes out his cell phone and reports your license plate and action to the local police.

There’s no question that the pedestrian is holding you accountable for failing to take sufficient care in driving, thereby placing him in danger. But he is also holding you morally responsible and the fact that he is doing that explains why his behavioral reactions toward you count as holding accountable. To see this, consider that you can make an angry face at someone just to be mean, even if you are not taking her to be
morally responsible for anything. Further, you can threaten someone if you want him to leave your vicinity, even if he hasn’t acted in any way at all toward you. You can remonstrate someone for an act you know he didn’t commit in order to try to decrease his favor with his supervisors. And, finally, you can prevent someone from enjoying some benefits he has a right to even though you don’t take her to have done anything wrong. Obviously, you can even behave toward someone in all four ways indicative of holding someone accountable at the very same time without taking her to be morally responsible. Since we can, and do, demonstrate these behaviors toward people who we do not hold accountable, why should we think of these reactions as canonical ways of holding someone accountable?

The answer concerns the fact that we think of these behaviors as ways of holding someone accountable because these behaviors are the characteristic\textsuperscript{121} behavioral upshots of holding responsibility-tracking attitudes. For instance, if our account of holding someone morally responsible were given purely in terms of nonemotional attitudes like trust, we would lack an understanding of why merely expressing anger toward someone can be a way of holding her accountable. You can come to distrust someone because she has acted with ill will toward you and therefore refuse to entrust things of importance to her without becoming angry with her. When you do so, you clearly hold her accountable for acting toward you with ill will. However, the attitude of distrust does not explain why responding to her act of ill will with anger would count as holding her accountable.

Alternatively, consider our central responsibility-tracking emotion: anger. Suppose you

\textsuperscript{121} In the next chapter, I have more to say about why we take certain behaviors to be characteristic of certain kinds of attitudes.
become angry with someone because her action showed ill will or insufficient concern, express your anger and reproach her. Again, you clearly hold her person accountable for what she has done. However, your anger does not directly explain why you might later fail to entrust her with things that matter to you.

So the fact that these behavioral responses are the normal upshots of the attitudes implicated in holding someone morally responsible explains why we take them to be ways of holding people accountable. But the relation between the moral responsibility-tracking attitudes and the accountability reactions is not just one of constant conjunction. It is not only *normal* to hold someone accountable when you hold her morally responsible, but we generally think that holding someone morally responsible makes holding her accountable more *reasonable*.

In saying that holding her responsible makes it reasonable to hold her accountable, I mean to point out that when you have an attitude that is a part of holding someone morally responsible, you often take yourself to have some reason to engage in a corresponding behavior that amounts to holding her accountable. So, for example, since the pedestrian is angry with you, it is likely to seem reasonable to him to express this anger to you. Further, since he believes that your driving poses a threat to him, he thinks that he has some reason to act in a hostile manner toward you on the assumption that threatening harm would prevent you from engaging in that manner of driving in the future. Since he wants you to understand the reason why he is expressing his anger and engaging in hostile behavior toward you, he takes himself to have a reason to berate you. Finally, because he finds you untrustworthy, he believes he has reason to alert the authorities to your untrustworthy nature so you can be further monitored.
Of course, whether these responses are reasonable, in the sense of justified, requires more argument than I have so far considered. But at this point, I want to emphasize that such argument could concern whether any attitude characteristic of holding someone morally responsible makes any part of holding someone accountable justified. It could be that fitting anger legitimates expressions of anger or hostile behavior. Fitting distrust might legitimate reproach or alerting the authorities. A full picture of our practices of holding people morally responsible and accountable will thus need to focus on both emotions, nonemotions, and their characteristic behavioral upshots if it is to determine when it is legitimate to hold someone accountable. It is understandable that much work on moral responsibility focuses on the reactive emotions and the associated behaviors involved in holding people accountable.

When we hold friends and relatives morally responsible, our extant relationships with them typically blunt the more unsociable elements in our accountability behaviors, so the attitudes and responses with which we are most familiar from our daily interactions are forms of anger and its behavioral upshots. It is therefore understandable that Pereboom’s primary question about Plum’s moral responsibility is whether he deserves blame—or, as I would urge, whether anger befits him. But focusing only on these emotional responses not only misses an important part of the picture about Plum’s moral responsibility, it also leaves out a centrally important issue for Plum’s accountability: whether his action gives us reason to refuse to entrust him with our welfare. I believe that this issue is of extended interest with respect to one of the so far neglected elements of our accountability behaviors: punishment. In the next chapter, I will explore more
carefully the relation between the attitudes of anger and distrust and the justifiability of punishment.
CHAPTER 5—PROFESSOR PLUM’S PUNISHMENT

5.1 INTRODUCTION

In the previous chapter, we saw that Professor Plum is morally responsible on a reactive nonemotion account of moral responsibility, which, coupled with the fact that he is morally responsible on a reactive emotion account of moral responsibility, means he is morally responsible on both reactive emotion and reactive nonemotion accounts of moral responsibility. This means the global manipulation argument fails to show that Plum is not morally responsible for killing White. Because Pereboom’s Plum case is the most impressive manipulation case currently on offer, my argument makes it plausible to think that compatibilism is immune to the manipulation argument. While this is a significant result, we can imagine an objection that the discussion has proceeded in a relatively rarified fashion. An objector might note that questions about someone’s moral responsibility are often broached when another topic is really at issue: namely, whether or not it is appropriate to punish him. While acknowledging that my arguments suggest\(^\text{122}\) that people can be morally responsible even in a deterministic world, the objector might urge that I have left unconsidered the crucially important question of whether it is

\(^{122}\) I say “suggest” rather than “show” since I have not considered every possible argument that purports to show that moral responsibility and determinism are incompatible.
appropriate to *punish* Plum. This is especially of concern because in Chapter 2, I allowed potential objectors to claim that Plum does not deserve anger. If I am willing to grant this, am I therefore required to hold that Plum does not deserve punishment and we are therefore not morally justified in punishing him? This would be worrying at a theoretical level as well, for the fact that someone is morally responsible for a certain kind of offense is typically thought to play a crucial role in explaining why we are morally justified in punishing her. What does my account of moral responsibility say about the justification of punishment for people who commit serious crimes as the result of being globally manipulated?

This chapter aims to address these issues. Let me first introduce some desiderata for a theory of the moral justification of punishment, which I will use to guide the discussion below. Such a theory must explain the way in which punishment is *offender-focused*; that is, it must make clear why it is only permissible to punish those who are morally responsible for committing an offense. A successful theory of the justification of punishment must also account for why it is essential to our conception of punishment that punishing an offender expresses *condemnation*.\(^{123}\) I believe the recent ascendance of retributivism in the philosophical literature on punishment is due to its seeming promise to explain why just punishment is offender-focused and why punishment expresses

\[\text{\footnotesize\(^{123}\) The definition of punishment as essentially involving both harm to and condemnation of an offender, is widely accepted. See, for example, Kelly and Bedau, “Punishment”; Dolinko, “Some Thoughts About Retributivism”; Fischer, “Punishment and Desert”; Hart, “The Aims of the Criminal Law”; Feinberg, “The Expressive Function of Punishment.”} \]
condemnation. I use “retributivism” to refer to any theory of punishment that attempts to explain the moral justification of punishment by appealing to the idea that criminals deserve punishment. Defining retributivism this way explains why retributive punishment would be thought to account for punishment’s offender-focused and expressive qualities; it seems clear that the guilty deserve punishment and that punishment aimed at giving people what they deserve expresses condemnation of their action.

Even though it might seem that retributivism is the only theory of punishment that could possibly fulfill both these desiderata, I believe there are other theories that allow us to avoid retributivism while still accounting for punishment’s offender-focused and condemnatory qualities. In this chapter, I want to develop the theoretical options open to someone who rejects retributivism but thinks punishment is offender-focused and expresses condemnation. Many who reject retributivism would likely seek a theory of the justification of punishment that emphasizes consequentialist aims like incapacitation or deterrence. However, not only would such theories seem to run afoul of the two

---

124 For a discussion of retributivism’s ascendance, see Dolinko, “Some Thoughts About Retributivism.”


126 There are good reasons to avoid retributivism independent of how it treats Plum’s case and there is no better source for reasons to reject retributivism than David Dolinko’s trenchant critiques. “Some Thoughts About Retributivism”; “Three Mistakes of Retributivism.”
desiderata above, such theories must confront the objection that punishment aimed at these ends unjustifiably uses offenders as a mere means.\textsuperscript{127} I believe it is possible to meet this objection and explain why we may permissibly engage in punishment that aims at incapacitation or deterrence without unjustifiably “using” offenders as a mere means when engaged in these pursuits. I also show that such justifications meet the two desiderata above and argue that the pursuits of deterrence and incapacitation can be given a nonconsequentialist,\textsuperscript{128} nonretributive\textsuperscript{129} justification that explains why we are justified in socially isolating Professor Plum.

My talk of ‘social isolation’ deserves a brief remark. While for the most part I do not explicitly consider arguments that bear on the justifiability of state punishment for Plum, I argue for a justification for socially isolating Plum that I believe bears important normative similarities to incarcerating him by a state authority. If I can show that socially isolating Plum is justifiable, then I will have demonstrated that we are justified in doing something very much like incarcerating him, since I believe that, with few modifications, the social-isolation theory can be adapted into a theory of criminal punishment.

\textsuperscript{127} The classic statement of this “use” objection is Kant’s, in \textit{The Metaphysical Elements of Justice}, trans. John Ladd (Indianapolis: Bobbs-Merrill, 1965), 100-101.

\textsuperscript{128} I use “nonconsequentialist” here to refer to theories that take the justification of punishment to be constrained in pursuit of beneficial consequences—like deterrence or incapacitation—by some other factor. On this terminology, a view that takes it only to be permissible to engage in deterrent or incapacitory punishment when someone is morally responsible for an offense will count as nonconsequentialist. Retributivism will also count as a nonconsequentialist view on this terminology.

\textsuperscript{129} By “nonretributive,” I indicate theories that do not take the moral justification of punishment to involve an offender’s desert.
My argument runs as follows: I first examine Derk Pereboom\textsuperscript{130} and Ferdinand Schoeman’s\textsuperscript{131} purely consequentialist justification for socially isolating those who have committed serious offenses that proceeds via an analogy with quarantine. I reject this account as justifying our practice of punishment because it fails to take seriously the fact that the moral responsibility of offenders plays a central role in determining the propriety of punishment. This defect makes it unable to account for punishment’s condemnatory character. While it might be thought that the only way for the moral responsibility of offenders to constrain the justifiability of socially isolating punishment is by adopting retributivism, I show this is not the case by briefly laying out a novel nonconsequentialist justification for the harming of offenders based on the work of Daniel Farrell.\textsuperscript{132} Farrell mobilizes his theory to show that we sometimes permissibly harm wrongdoers in the service of both special and general deterrence. I show that the theoretical framework that Farrell mobilizes also applies social isolation aimed at incapacitation and so the theory demonstrates that we would be permitted to socially isolate Plum. However, just like the quarantine view, the account I develop also faces the objection that it would not justify \textit{punishment} but merely activities of social management, since punishment essentially involves condemnation of the offender and the account I develop does not articulate a sense in which social isolation aimed at deterrence or incapacitation essentially involves condemnation. To answer this objection, I explain that when the nonconsequentialist theory justifies social isolation, it does so only when anger and distrust are fitting toward

\textsuperscript{130} \textit{Living Without Free Will}, 174-177.
\textsuperscript{131} Ferdinand D. Schoeman, “On Incapacitating the Dangerous.”
\textsuperscript{132} “The Justification of General Deterrence”; “Punishment Without The State”; “The Justification of Deterrent Violence.”

114
the offender. I argue that if we structure a practice of social isolation in this way, we can see that the nonconsequentialist theory captures punishment’s condemnatory aspect.

Finally, I show that my account explains why we may well be justified in punishing Plum while still respecting the intuition that Plum does not deserve punishment because of the actions of the neuroscientists. It can thus account for a reluctance to punish him because he was manipulated, even though socially isolating him may be permissible. Interestingly, then, my account suggests that desert plays no role in explaining the moral justification of the social isolation. This allows the nonconsequentialist theory to account for conflicting intuitions about punishing Plum.

5.2 Purely Consequentialist Incapacitation

I begin by discussing Derk Pereboom and Ferdinand Schoeman’s purely consequentialist justifications for the social isolation of those who commit serious crimes, which proceed via an analogy with quarantine. Though I believe their accounts fail to properly conceptualize the social isolation for which Plum is eligible, I start here for two reasons: first, because their accounts are two of the only discussions in the philosophical literature that specifically discusses socially isolating treatment—the primary form of punishment used in Western democracies—and second, because this is Pereboom’s favored way of dealing with people who have committed serious crimes. Pereboom follows Schoeman in arguing that if we take ourselves to be justified in confining people with dangerous communicable diseases, we should similarly think ourselves justified in engaging in civil detention of criminals. After all, if the danger to society is large enough, many of us think that we are justified in confining those who have serious
communicable diseases. Importantly, it seems that our justification for doing so is not essentially connected to the moral responsibility of the carriers of the diseases. Justified quarantine does not require that the people who are ill be morally responsible for the fact that they pose a danger to the community in virtue of having the disease. If this is correct, then with the sort of view defended by Pereboom and Schoeman, we get a defense of socially isolating incapacitation that does not require moral responsibility on the part of the offender. Still, in spite of its attractiveness, their theory is strained in application to criminal punishment.

The primary reason why is that the theory fails to respect the fact that criminal punishment does not just serve to achieve social goals like deterring potential criminals or isolating convicted ones. Punishment does often aim at these positive social consequences, of course. But we fail to properly conceptualize punishment as such unless we conceive of it as not only doling out harms in order to make us safer by offering disincentives that will modify the behavior of people in a society, but as essentially involving an expression of condemnation of the offender’s action. Pereboom and Schoeman’s view cannot make sense of the condemnatory aspect of punishment. To see why, consider that the primary message sent by a punishment system structured as recommended on Pereboom and Schoeman’s view is merely that it is dangerous for certain people to freely interact with other people. Nothing in the justification of quarantine makes any reference to the fact that quarantine is justified specifically in relation to people who are morally responsible for having done things that are seriously wrong or unjust. This is problematic, and not simply because the quarantine view fails to respect punishment’s condemnatory aspect. Because it does not essentially link
punishment to actions it is fitting to condemn, the quarantine view would sometimes advocate punishing innocents if it were treated as a theory of the justification of punishment. As Phillip Montague notes when critiquing a purely consequentialist theory of punishment,

a critic of act utilitarianism for punishment might grant that punishing innocent people can be justified in special circumstances, while continuing to maintain that even in these circumstances the innocence of those who are punished creates a moral presumption against punishing them—a presumption that act utilitarianism for punishment cannot accommodate. 133

The same critical point is true of the quarantine view of punishment, which does not treat an offender’s moral responsibility for doing wrong and the lack of moral responsibility of nonoffenders as having any role, much less an essential one, in structuring our practices of criminal incarceration. This is a fundamental mistake, of which the failure to ground the presumption against punishing the innocent is just one aspect.

The larger problem for the quarantine theory, and other purely consequentialist theories of the justification of punishment, is that it does not take seriously the fact that punishment tries to morally prioritize the innocent and protect them from danger posed by offenders. The quarantine view simply attempts to prevent harms from befalling members of the population by erecting barriers to the transmission of harms. If we imagine a situation in which the healthy are at risk of harm but in which there are far more sick people, the quarantine view will necessarily recommend restricting the liberty of healthy people. In its application to criminal punishment, then, the same thing would be true in a situation in which a large population of criminals preys on a few innocents. Suppose, for example, the neuroscientists create a small army of Professor Plums who

prey on the members of a particular community. If setting up barriers to the freedom of movement of the innocent requires fewer resources and the restriction of a smaller number of people’s liberty, and has results that are better than the results of restricting the movement of the Plums, the quarantine view actually recommends restricting the liberty of the innocent rather than the army of Plums.

The quarantine theory simply attempts to prevent harms from befalling members of the population by erecting barriers to the transmission of harms. It thus does not take seriously the fact that punishment tries to morally prioritize the innocent and protect them from danger posed by offenders. Because it does not do so, it also cannot account for punishment’s condemnatory aspect. Such a result is at odds with the offender-focused nature of punishment, so those of us searching for a plausible nonretributivist justification for punishment must look elsewhere.

5.3 Harm Justified in the Pursuit of Deterrence

In order to begin to develop a nonconsequentialist account of punishment that takes punishment’s connection with guilt and condemnation seriously, I would like to move our attention from the context of institutional, state punishment to consideration of interpersonal responses to wrongdoing in a Lockeian state of nature.\textsuperscript{134} I believe this change of focus is helpful, because I doubt very much that a political society would be justified in punishing criminals unless it were also true that there were interpersonal rights to punish in a state of nature (or, at least, rights to do something that looks very

\textsuperscript{134} Readers familiar with Daniel Farrell’s work on punishment will note parallels to his argumentative strategy in “The Justification of General Deterrence”; “Punishment Without The State”; “The Justification of Deterrent Violence.”
much like state punishment and is relevantly similar to state punishment in certain ways). Further, if such a right does exist in a state of nature, individuals could presumably transfer that right to states, thereby giving states the right to punish. Therefore, considering our rights in a Lockean state of nature is a good proxy for thinking about the rights that states might have over individuals in a political society.

In this spirit, I now want to turn to Daniel Farrell’s work on the justification of deterrent violence in such a Lockean state of nature. Farrell’s defense of such violence toward offenders proceeds by appealing to the following principle of justice that Farrell claims grounds our intuitions about justified self-defense:

> When someone knowingly brings it about, through his own wrongful conduct, that someone else must choose either to harm him or to be harmed herself, justice allows the latter to choose that the former shall be harmed, rather than that she shall be harmed, at least if the harm inflicted on the former is roughly proportional to the harm that would otherwise be inflicted on the latter.

Specifically, Farrell’s principle grounds the following self-defensive rights. The first is the right of direct self-defense: the right to resist, with deadly force if necessary, another person’s unjust attempt on my life. Someone who embraces the idea that each of us

---

135 To see why, consider that we would not think that the expressed preferences of a majority (or super-majority) of citizens are enough to show that a state may legitimately criminalize just any sort of act. In order for an act to be legitimately criminalized, the state must have a legitimate interest in preventing it: for example, the act must be a threat to the welfare, liberty or property of other citizens. Acts that do not fall into one of these categories (or other areas where the state has a legitimate interest) may not permissibly be criminalized. This suggests that there are moral barriers to punishment that are conceptually prior to the existence and legitimacy of states; no state may legitimately violate these barriers in punishing.

136 Farrell’s view develops over the course of the three articles referenced above, but always relies on a principle of justice such as the one cited. This version of the principle is from “The Justification of Deterrent Violence,” n.d., 302-3, http://www.jstor.org/stable/2380998.

137 “The Justification of General Deterrence.”
possesses the right of direct self-defense can be led to realize that we each also hold another related right, which Farrell terms the right of *indirect* self-defense. This is the right to threaten a previous attacker with serious physical harm to keep him from attacking me in the future and then, if necessary, to actually harm him in order to make the threat effective, if doing so is the only way to prevent him from attacking me again.

If Farrell is correct that his principle grounds the rights of direct and indirect self-defense—and I think he is—then there are interesting and obvious parallels with punishment for the purposes of special deterrence: punishment aimed at deterring the offender himself. Farrell’s argument is able to show that the pursuit of special deterrence, which is often thought of as a purely consequentialist justification of punishment, can be suitably “offender-focused”; that is, his theory can explain why punishment for the purposes of special deterrence is justified when and only when it aims at harming someone who has already seriously wronged others. But Farrell does not stop there; he also applies his insights to the justification of general deterrence: punishment aimed at deterring other potential offenders.

Farrell notes that some attacks, or perhaps some sorts of attack, plausibly have more than just *immediate* harmful effects on us. Imagine, for example, that because of the success of a previous attack, you are more vulnerable to future attacks by other aggressors. (This might be because the previous attack left you weaker than you were before, or because potential aggressors are emboldened after seeing that you have been successfully attacked.) But now suppose that by harming the offender who previously attacked, you could counter the additional vulnerability by giving other potential aggressors pause. If that were the case, then by the principle of distributive justice
introduced above, you would be justified in harming the initial offender, other things being equal, because his previous wrongful action forced the choice between harming him or being subject to wrongful harm from him. Note that this last appeal to the principle serves to emphasize that it circumscribes permitted harms and it implies the existence of certain limits on the harms to which offenders are liable. Specifically, offenders are liable to additional harms only if the additional harms applied to the offender will prevent vulnerabilities to innocents that might result from the offender’s action. Further, only harms that are proportionate to the harm that would otherwise be inflicted by the offender are justified; Farrell’s view does not defend the idea that just any harm applied to an offender for the sake of general deterrence is justified.

How would Farrell’s view address Plum’s case? Farrell’s view would allow us to harm Plum, so long as doing so would counter vulnerabilities that other innocents might suffer. For example, suppose that Plum’s killing Ms. White would lead to four additional unjust killings of other innocents because potential attackers were emboldened by the ease with which Plum slew White. However, if we could warn off those potential

---

138 In Farrell’s most recent paper, he considers whether the justification he offers for deterrent harm could justify a system of threats and harms meant to lower the absolute probability of unjust harm to the innocent by raising the conditional probability that those who unjustly harm the innocent will be harmed. His conclusion is that it would, which would therefore seem to justify some harms to offenders that do not prevent any innocents from becoming more susceptible to harm. However, as it is only at the level of a system of threats that we might be justified in harming offenders when the harm will not prevent vulnerabilities to harm that result from the offender’s own action, I ignore this issue to simplify the discussion. “The Justification of Deterrent Violence.”

139 Note that the structure of Farrell’s view might allow it to justify quite a bit more than simply social isolation, for his view is about the justification of harm, generally, not simply socially isolating harms. I set aside this issue to focus on the justifiability of socially isolating offenders because if his view justifies more serious harm, it seems clear that it justifies social isolation.
attackers and thereby prevent the four unjust killings by harming Plum, then by appealing to Farrell’s view, we can show that we would be justified in isolating Plum to the additional degree necessary to warn off the potential attackers. Further, we would be justified in harming Plum if doing so would be effective at deterring him from killing in the future.

5.4 BEYOND DETERRENT VIOLENCE

Farrell’s theory points us in a very helpful direction for understanding the moral justification of punishment, but his discussion of the implications of his view does not show us all the moral parallels that can be drawn between individual self-defense and state punishment. In this section, I’ll show that Farrell’s account has application beyond his original discussion and that being explicit about this allows us to see his theory as a comprehensive moral justification of punishment.

The first way in which I want to insist that Farrell’s view has additional application is to explicitly emphasize that the harms to offenders that the view justifies are not limited to physically damaging harms. I believe some philosophers have been led astray by the fact that Farrell sometimes speaks of his project as involving an investigation of the justification for inflicting pain on offenders,140 while other times he speaks of justifying the infliction of suffering on offenders141 and also of the justification of deterrent violence.142 This has led commentators like Pereboom to object to Farrell’s theory on the grounds that it attempts to justify the infliction of pain and suffering once a

140 “The Justification of General Deterrence,” 369-70.
141 “Punishment Without The State,” 443.
142 “The Justification of Deterrent Violence.”
criminal is already in custody. Surely if a criminal were already in custody, Pereboom urges, only “isolating the criminal from those he would endanger”\textsuperscript{143} could be justified. Therefore, Pereboom takes Farrell’s theory to generate an unpalatable result in such a case.

It should be clear that Pereboom’s objection reflects a misunderstanding of Farrell’s view. On Farrell’s account, you are justified in defending yourself—both in direct and indirect self-defense cases—only with the minimum amount of harm necessary to prevent harms that would result from an offender’s action (so long as the harm you inflict is not in excess of the proportionality limitation.) If, because an offender is already in custody and the harm done to him by keeping him in custody is sufficient to prevent the future harm to you and others that would have been engendered by his wrongful action, then there is no need to harm him further to prevent future harms. Farrell’s view thus provides no justification for inflicting pain and suffering on an offender that does not serve the aim of deterring harms that would result from the offender’s action absent intervention.

Even though Pereboom appears to misunderstand the implications of Farrell’s account, Pereboom’s misunderstanding is an opportunity for increased theoretical clarity. Since Farrell’s discussion typically emphasizes warding off a physical harm with another, it is possible to miss the fact that the principle of justice Farrell discusses could be applied to noncorporeal harms—like social isolation. But socially isolating someone

\textsuperscript{143} \textit{Living Without Free Will}, 174. Pereboom appears to ignore the possibility that by additionally harming the offender, you might successfully deter the offender or others. If such harms were engaged in pursuant to Farrell’s principle of justice, it’s not clear what would be problematic for Farrell’s view.
strikes Pereboom as fundamentally distinct from what punishment is about. Pereboom continues:

It is not obvious that theories like Farrell’s can justify punishment of criminals, given that punishment involves the intentional infliction of significant harm, such as death or severe physical or psychological pain, as opposed to, for example, “quarantining” them.\(^{144}\)

While there is no doubt that *some* forms of punishment, such as the death penalty, do involve the application of significant bodily harm, it is a mistake to think of sanctions like incarceration as something other than *punishment* because they do not involve the infliction of bodily harm. Incarceration does involve a very severe circumscription of a person’s liberty, which *is* a significant harm, even though it does not involve bodily injury.

There is a further issue regarding Farrell’s presentation of his view that is also of some concern. When Farrell lays out the implications of his principle of justice, he only examines how it justifies harms to offenders when those harms successfully prevent future harms to innocents by *deterring* offenders from engaging in harmful activity. But there are surely cases in which harm to an offender—specifically a harm like social isolation—prevents harm to innocents without doing it through deterrence and Farrell’s theory straightforwardly applies to such cases.

To see this, return for a moment to the application of Farrell’s account to Plum’s case. Showing that harming Plum could be justified by Farrell’s theory involved some speculation about future harms that might result from Plum’s action. But suppose that no future harms would result from Plum’s action except for those that Plum, himself, might

\(^{144}\) Ibid.

124
perpetrate. And, further, suppose that Plum would not be deterred by the threat of harm from performing additional wrongful acts. If Farrell’s theory could only justify harming someone in the pursuit of deterrence, it would seem to have an unpalatable result: it would not justify harming Plum, as threatening him with harm could not successfully deter him. However, if Plum could not be deterred from performing future wrongful acts, we would certainly want to protect ourselves from him. In the next section, I will show that Farrell’s theory is not restricted to social isolation aimed at deterrence, but also provides moral justification for incapacitive social isolation.

5.5 INCAPACITIVE SOCIAL ISOLATION

Suppose we’re back in the state of nature. Consider, in parallel to the rights of direct and indirect self-defense discussed above, what I call the rights of direct and fiduciary social isolation. To characterize the right of direct social isolation, suppose that a person who lives nearby killed my neighbor and that nothing my neighbor did made such an attack justifiable. Regardless of any other rights we might have in relation to this person (whom I will term the “offender”), most of us would say that I have the right to take certain steps to prevent the offender from being able to successfully attack me in the future. For example, most of us would say that I am justified in building a fence that will effectively keep him off my property and that I am justified in traveling with a group of guards who will be able to keep him a safe distance away from me. Further, if the only way to keep him from attacking me in the future is to confine him to his property, at least some of us would say that is permissible. These actions would be considered examples of direct social isolation.
Besides the right of direct social isolation, I believe that many of us would also claim for ourselves a right of *fiduciary* social isolation. This is the right to take similar sorts of steps with respect to the offender to ensure that he is unable to attack someone else: my child, or perhaps one of my friends. If an offender had recently killed our neighbor, many of us would say we have the right to take the defensive and protective steps outlined above on behalf of others, for example, by putting up a fence around our child’s play-area. If we had the means, we would take ourselves to have the right to employ guards to keep the potential attacker away from our friends. And we might think ourselves permitted to confine the offender to his property if that was the only way to keep the offender from harming additional innocents.

Just as with the right to harm offenders in the pursuit of deterrence, our exercise of the rights of direct and fiduciary social isolation are properly governed by Farrell’s principle of justice. Recall the principle:

> When someone knowingly brings it about, through his own wrongful conduct, that someone else must choose either to harm him or to be harmed herself, justice allows the latter to choose that the former shall be harmed, rather than that she shall be harmed, at least if the harm inflicted on the former is roughly proportional to the harm that would otherwise be inflicted on the latter.\(^{145}\)

We can see that this principle applies to the exercise of the rights of direct and fiduciary social isolation by first noting that socially isolating someone is generally a harm to him. Social isolation, as I construe it, amounts to refusing to allow a person the ability to enter, or have control over, a social domain. Such isolation could be relatively minor (not being willing to enter into a certain set of agreements with the person) or could be quite major.

---

(taking steps to prevent the person from having any social contact with me, my fiduciaries, and our property). Since socially isolating someone constitutes a harm to that person, the principle of justice above only justifies such harms when (1) someone knowingly brings it about through his wrongful conduct that someone else must choose either to socially isolate him or to be harmed herself, and (2) the harm inflicted via social isolation on the former is roughly proportional to the harm that would otherwise be inflicted on the latter. In the remainder of this section, I will demonstrate that (1) and (2) apply to the exercise of the rights of direct and fiduciary social isolation.

Someone might doubt that (1) applies to a situation where someone has already knowingly taken a seriously wrongful action. After all, the objection would go, the harm that action has caused is completely in the past—there is no additional harm to counteract with a socially isolating action. I believe this is often false, though below I consider the implications for the view when it is true. To see why it is often false, consider a Lockean state of nature situation where all parties in the state of nature coexist peacefully. Then one of the parties knowingly takes an action that seriously wrongs another person. Suppose the immediate wrong is a physical harm. While that harm is in the past and cannot be directly counteracted with social isolation of either type, there is another ongoing harm to the person who was wronged, as well as to other people suitably related to the victim. That harm is the risk of future harm at the hands of the offender.

---

146 My argument need not assume that property holdings can be justified in a state of nature. The concept of a social domain concerns all possible social interactions between persons. Property relations are merely one of many possible forms of such interactions. If property cannot legitimately or coherently be held in a state of nature, my account can proceed by appeal to other forms of social interaction.
Now, an objector might reply that it is quite a stretch to call a risk of future harm an actual *harm*. After all, this is not the direct or immediate harm that someone faces in a self-defensive situation. This is correct. Note, however, that even in such an immediate self-defense situation, the defender is typically responding to a mere risk of harm (however likely). It is not as if the offender brings a harm into the world which the defender, after perceiving it as fully extant, simply redirects back onto the offender. Rather, even in a direct self-defense situation, the attacker presents a *threat of or attempt* to harm, which, while very *probable* (absent self-defensive action), has not yet been successful. Someone’s attack on you gives you very good evidence that if you do not act, you will be harmed, but the harm does not yet exist. My contention is that in a similar manner, the past wrongful action of an offender constitutes evidence that the offender might engage in future wrongful actions. When you have such evidence, to continue to engage in normal social interactions with the wrongdoer exposes you to an increased risk of wrongful harm and that risk should be treated as a harm by the lights of the principle of justice. As Farrell argues,

[W]hy should we balk at calling this [increased risk of wrongful harm] "harm"? Someone who is suddenly put at greater risk of wrongful harm is certainly worse off than someone whose prospects have not been affected in this way, and being put in such a situation is something that most of us, at any rate, would readily take great pains to avoid.¹⁴⁷

It is plausible, then, that the ongoing threat posed by a person who has knowingly harmed someone via a seriously wrong action is a harm, or, at least, should properly be treated as such for the purposes of being addressed under the principle of justice above.

¹⁴⁷ “Punishment Without The State,” 444.
However, an objector might reply that while the ongoing threat posed by the offender does present a risk of harm, it is hardly a risk that is created by the wrongful action of the offender. After all, the objector might urge, surely the risk of harm at the hands of the offender was there all along—the offender’s wrongful action was an instance of the general threat posed—and so is not eligible for treatment by our principle of justice. This observation about the risk being already present is correct. The offender’s wrongful action, itself, does not create the risk of harm. The offender’s action instead brings it about that we must choose whether to allow that risk of harm to continue, or else to eliminate the harm by socially isolating the offender. The principle applies not because the offenders’ action brings about the risk of harm, but because the offender’s action makes someone choose that risk of harm to the innocent or a harm to the offender.

Consider, by way of comparison, one of two sorts of harms produced by the wrongful actions of offenders that Farrell treats as eligible to be prevented by deterrent violence governed by the principle of justice. One of these harms is the increased likelihood of harm at the hands of other potential offenders who are emboldened by the success of the current offender. But note that the reason we may choose between the risk of harm by other potential offenders and a sure harm to the current offender might be that the wrongful action of the current offender has educated us—and potential offenders—about certain risks or wrongful harm to which innocents are already currently exposed. Other potential offenders might already be willing to perpetrate their own harms and have the desire to do so; they may simply be casting about for information regarding where innocents can easily be preyed upon. In such a situation, there need not be any new risks
created by the wrongful actions of the current offender. Innocents are already at risk from the actions of other potential offenders, but the risk to innocents is broadcast as the result of the first offender’s action. When the information about risks becomes public, that information may be acted on by both potential offenders and those who are motivated to protect the innocent from additional wrongful harm. Information about the offender’s dangerousness is produced by his past wrongful action and that makes us realize we face the choice between harming him or accepting the risk that emboldened potential offenders might harm innocents.

So in making it known that certain sorts of innocents or their property are vulnerable to wrongful attack and that others might take advantage of that vulnerability, the offender does make it the case that someone must choose either to harm him or else accept an increased risk of harm to innocents. He does so not by creating a new harm, but rather by creating a new opportunity for choice. This is also true with respect to the exercise of the rights of direct and fiduciary social isolation governed by the principle of justice. An offender who knowingly commits a seriously wrongful and harmful action does bring it about that we have to choose between two harms: the risk of his harming the innocent or a harm to the offender via social isolation. This is a choice that we didn’t know we faced before the wrongful action of the offender. His action demonstrates we are in that situation because his wrongful action constitutes evidence that he is dangerous—evidence that we lacked before he acted.

To this point I have argued that the decision about whether to exercise the rights of direct and fiduciary social isolation is a decision about how to distribute harms via Farrell’s principle of justice, just as it is in the decision to exercise the rights of direct and
indirect self-defense. It is true that one of the harms is not an actually existing harm, instead representing the probability of future harm, but again, this is also true in cases of direct and indirect self-defense. Further, I have argued that the offender’s conduct has brought it about that because of his wrongful conduct, someone must choose between harming him or accepting the risk of harm to the innocent. At this point, an objector might reply that if we encounter a situation where somehow we know with certainty that a person who previously perpetrated a seriously wrongful harm against someone will not ever commit such a offense again, then the framework I have sketched so far will not apply. I grant this. But this is no objection to the view. For if we knew with certainty that an offender would not reoffend and that socially isolating him would not prevent other harms to innocents though deterrence or incapacitation, it is very hard to see why we would want to punish him, unless, of course, we wanted to do so for the sort of retributivist reasons I reject.

The remaining issue that I need to address is the issue of proportionality. Recall that the principle of justice to which we are appealing only allows us to inflict harms in the service of preventing harms wrongfully perpetrated by others if the harms we inflict are proportionate to the harms we seek to prevent. But what would it mean to say that a certain risk of harm is proportionate to a certain actual and direct harm? That is, what would it mean to say that the risk of future harm from an offender’s future action is proportionate to the harm of the social isolation we inflict on him?
As Farrell notes, there are two distinct issues here.\textsuperscript{148} The first is the issue of clarifying what we \textit{mean} when we say that certain harms we might inflict on offenders are proportionate to certain risks of harm we seek to prevent. That issue can be fairly quickly dealt with, for what we mean in this context is the same sort of thing we mean when we say that one actual harm is proportionate to another. That is, a given harm is proportionate to a given risk of harm if the former is no more than what we are entitled to do to someone as a way of avoiding the latter.

But that schema doesn’t tell us the answer to what we are actually \textit{entitled} to do to someone, even if we are attempting to keep the harm of the social isolation we impose proportionate to the risk of harm we seek to avoid. As Farrell notes, there are essentially two categories of answer here. One sort of answer we might give is that the fact that we are balancing risks of harm against actual harms is irrelevant. If we adopt this sort of answer, then we would say that if the risk of harm from an offender’s future action is death, then we would be entitled to completely and irrevocably socially isolate him in response. Call this the \textit{extreme} view about proportionality. The \textit{moderate} view, in contrast, holds that the fact that we are balancing the actual harm in social isolation against a risk of harm from an offender’s future wrongful action is essential to what we are justified in doing to the offender. On the moderate view, then, the social isolation we are entitled to do to an offender is given by a probability-weighted function of the harm that the innocent might suffer if we do not socially isolate the offender.\textsuperscript{149}

\textsuperscript{148} Ibid., 445-7. My argument addressing these two points is heavily indebted to Farrell’s discussion.
\textsuperscript{149} Another relevant variable concerns the degree of culpability of the potential harmer. For example, it is plausible that if someone presents a harm to me because of her
I won’t try to work out the details of either approach in more detail here. The important point to note is that no matter which view we adopt, we will be socially isolating people in line with the above principle of justice. Thus, we will be socially isolating people in a way that, in contrast to the quarantine view, is offender-focused.

5.6 Socially Isolating Plum

At this point, my characterization of the view that social isolation is justified in relation to offenders whose past wrongful actions force a choice on us between risking further harms from them or socially isolating them is, admittedly, just a sketch. Much more would need to be said to fully defend it. Still, I believe this sketch is helpful, as it presents an offender-focused theory of incapacitive social isolation that is not retributivist. Above, we saw that if we made certain assumptions, we could apply Farrell’s deterrence-based theory to Plum’s case. I have argued that we possess a right against those who act in seriously wrongful and harmful ways that justifies us in socially isolating them to the degree necessary to prevent them from having control over the domains in which they risk harming innocents. At this point, it should be obvious that if Plum presents a future threat, the theory I have explored would hold that incapacitating Plum by socially isolating him is justified.

However, we can imagine an objector asking how we could possibly be justified in socially isolating *Plum* in response to Ms. White’s death. After all, he was controlled by the neuroscientists, who are ultimately responsible for Plum being willing to kill her negligent action, I am not justified in harming her as much as I would be if she presents a harm because of a purposeful action. For simplicity, I omit discussion of this additional variable.
for the sake of some small personal advantage. Shouldn’t we be socially isolating the neuroscientists?

I have no qualms about the justifiability of socially isolating the neuroscientists, as we will see in a moment. But consider first Professor Plum. Are we justified in isolating him even though his action was causally determined by the neuroscientists? If Plum—or someone else—objects to his social isolation, what sort of argument would he have to make on his behalf? The argument that would have to be made is that he is unlikely to reoffend—that in the relevant social domain he poses little risk to innocents through future wrongful action. In order to make this argument, an objector would have to demonstrate that there is little risk of Plum demonstrating such ill will again. But that argument is not bolstered by the fact that he was manipulated, so the fact that the neuroscientists manipulated Plum provides no basis on Farrell’s theory for objecting to Plum’s social isolation.

What, then, of the neuroscientists? Can the theory justify punishing them? It can. In fact, whatever amount of social isolation we determine is appropriate for Plum

---

150 We can imagine that we do have very strong reason to think that he will not reoffend, and in that case the framework will not justify socially isolating Plum. However, as I noted above, this is no problem, for if Plum will not reoffend, the only reasons to punish him would be retributive, which I have ruled out by assumption.

151 There might be arguments unrelated to Plum’s moral responsibility that could show that it is impermissible to socially isolate him. Suppose, for example, that the neuroscientists implanted a small bomb in Plum that could not be removed without killing several people and that would explode if he were incarcerated. In such a situation, we might be unjustified in punishing him, but not as the result of a lack of responsibility for performing the action for which we are considering punishing him. Such arguments therefore miss the point of the current discussion.

152 I’m assuming that the neuroscientists control Plum in the way that they do because they plan on causing White’s death via Plum’s action. But they might not be planning to have him kill anyone; they might simply be testing their mind control device’s efficacy.
is likely less than the level that would be appropriate for the neuroscientists. This is because the neuroscientists commit two very serious wrongs. They bring about Ms. White’s death via Professor Plum’s actions, and they also commit a serious wrong against Plum by controlling his mental processes so that he will kill White.\textsuperscript{153} If they might engage in such manipulation again, they give us even more reason to socially isolate them than Plum, because as he has committed only one serious offense, his actions pose less risk of harm to the innocent. As the two wrongs the neuroscientists perpetuate affect the probability of the future risk of harm to innocents, the theory likely justifies socially isolating them to a greater degree than we would be justified in socially isolating Plum.

To this point, we’ve briefly seen that two aims of punishment—incapacitation and deterrence—commonly thought to be supported by purely consequentialist justifications for punishment can be made suitably “offender focused” so that punishment in the service of those aims is not vulnerable to the objection that by engaging in such punishment we are “using” offenders in an objectionable way.\textsuperscript{154} The account on offer

\textsuperscript{153} David Sussman’s account of torture might provide an interesting parallel to the unique wrong in which the neuroscientists engage. I’m not assuming any particular account of why their control of Plum is wrong, however, just that it is a serious moral violation. “What’s Wrong with Torture?,” Philosophy & Public Affairs 33, no. 1 (2005): 1–33.

\textsuperscript{154} To fully defend these views, quite a bit more would need to be said, even apart from the condemnation issue. For example, the social isolation view needs to offer an account of how to determine the relevant social domain in which an offender is distrusted and we would need to say more about how Farrell’s view limits the amount of punishment in pursuit of general deterrence. As I believe these issues are tractable, I want to set them aside to focus on whether these views can account for punishment’s condemnatory aspect.
has justified engaging in social isolation toward offenders for incapacitive and deterrent reasons while not depending on any form of retributivism.

Still, an objector might respond that while in contrast to the quarantine view, these views do link the moral responsibility of offenders to the justifiability of isolating them, it is not clear that they can explain why condemnation is an essential feature of punishment. To see why this might be a concern, imagine an objector pointing out that even if we were to take the distrust and deterrence views out of the state of nature and use them to structure a society’s approach to managing criminal behavior, they would not yet constitute punishment because socially isolating someone, in and of itself, need not express any condemnation of a person’s action. While admitting that incarcerative punishment is typically thought to express condemnation, this objector urges that social isolation does not have to express condemnation, as shown by the quarantine example.

However, punishment’s condemnatory character can be accounted for by the nonconsequentialist accounts I explore by appealing to my earlier discussion of fittingness. Explaining that the sanctions are fitting allows these views to account for punishment’s condemnatory force. I will argue that social isolation is fitting because the deterrence and distrust theories link the justification of punishment to the offender’s moral responsibility for committing unjust or seriously wrong acts. If my argument is successful, then I will have argued that these accounts do not merely offer techniques for justified social management, but that they justify something that deserves to be called “punishment.”

155 That is, something that would properly be called “punishment” if it were moved from the state of nature into civil society.
5.7 PUNISHMENT’S EXPRESSIVE FORCE

We saw above that punishment is generally thought to involve both harming and condemning offenders. It seems intuitive to some, including Pereboom, that punishing people by imposing corporal harms on them would express condemnation. But it is not so clear that simply socially isolating people in accord with the recommendations of Farrell’s principle would similarly express condemnation. After all, we can imagine that the officials authorized to administer the social isolation impose it completely dispassionately. Why, then, should it be thought to express condemnation of the offender’s action? I believe we can explain this phenomenon by appealing, once again, to anger—the emotion most naturally connected with punishment. I say anger is the emotion most naturally connected with punishment because in societies that do not have formal legal systems, anger and its associated expressions are the primary mechanism by which people ensure their rights and claims are respected.\(^{156}\) Anger also serves that function in our society when it would be in neither party’s interest to bring the mechanism of the formal law to bear on the dispute. So anger’s function is similar to that of an aim of the criminal law: to encourage people to respect important rights and claims of citizens and to call them to account when they do not.

However, punishment does not just bear functional similarities to anger. Anger’s conditions of fit are isomorphic to punishment’s conditions of justifiability. Recall anger’s conditions of fit are that someone performs an action that transgresses a socially significant norm because that person is motivated by ill will or lacks required good will,

and that situation is bad in some way. The conditions of justifiability of isolation on the social isolation theory are a subset of these conditions, namely, where the norm violated is seriously wrong or unjust. And the conditions of justifiability of isolation on the deterrence theory are a subset of that set of conditions, namely, where additional social isolation can prevent additional harms that might result from the offender’s action.

I also want to suggest that anger is not the only reactive attitude that bears important conceptual similarities to punishment—distrust is another. We have seen that some nonemotional attitudes—like trust and distrust—are also reactive attitudes. As distrust is also fitting when competent people show ill will or demonstrate insufficient concern, punishment’s conditions of justifiability are also a subset of distrust’s conditions of fit. Further, distrust’s social function is arguably to help encourage the avoidance of harms perpetrated by those who are untrustworthy, which is quite similar to the function of social isolation: preventing harm to the innocent by limiting the liberty of the untrustworthy.

Therefore, we understand the condemnatory aspect of punishment by appeal to attitudes that share with punishment similarity of functions, as well as an isomorphism between the conditions of fit of the attitude and the conditions of justifiability of the treatment. Let us apply the account to Plum to demonstrate that socially isolating him would express both anger and distrust toward him. Socially isolating Plum fulfills similar functions to becoming angry and distrustful of him and has conditions of justifiability that are isomorphic to the conditions of fit of anger and distrust. Because of this, we can say that social isolation applied to Plum on the basis of the deterrence and distrust views will be fitting for him—not because, as in the case of fitting attitudes, an evaluative
presentation is accurate, but because punishment’s conditions of permissible application bear the above similarities to the evaluative presentation of the attitudes that find their expression in punishment and other sanctioning activities. So for these reasons, socially isolating Plum expresses to him that we feel anger and distrust toward him.

To condemn someone’s behavior is to express disapproval of it. I have argued that socially isolating Plum in accord with Farrell’s principle of justice would express anger and distrust toward him. These attitudes are condemnatory attitudes because they are fitting in response a particular sort of content: the fact that someone’s action has shown ill will. So socially isolating him in accord with Farrell’s theory communicates to him that we view him as a person who, because he has shown past ill will, is subject to a scheme of social isolation aimed at preventing seriously wrong and unjust actions. Isolating him, then, communicates to him that we do not view him as willing to engage with us in fair and appropriate ways. It is this content—that Plum is not willing to engage with us on fair and morally appropriate terms—that is expressed to Plum in socially isolating him. Thus, taking actions that prevent Plum from having social contact with us because of his past ill will sends him the message that he is untrustworthy and unwilling to engage with others in ways that are not seriously wrong. So for these reasons, socially isolating Plum in accord with Farrell’s theory also condemns his killing of Ms. White.

We are also now in a position to explain away another challenge to my view. Since Pereboom was pushing the claim that Plum did not deserve anger, surely he would also advocate that because of the actions of the neuroscientists, Plum does not deserve punishment for killing Ms. White. Just as I granted Pereboom the claim that Plum does
not deserve anger in Chapter 2, I am in a position to grant the claim that he does not
deserve punishment as a way of accounting for the thought that it would be wrong to
punish Plum for killing White. If people react to Plum’s case with the thought that it
would be wrong to punish him, they are likely building from the thought that Plum would
not deserve punishment for killing White. But punishing Plum (or, at least, something
very like punishment: social isolation) can be fitting and just, even as it is undeserved.
The fact that Plum does not deserve punishment, then, does not show that we are not
permitted to punish him.

My account can also explain why it might seem so obvious to so many that
retributivism is correct as a moral justification for punishment. As we’ve seen on the
social isolation theory, harming offenders by isolating them is justified when offenders
have acted in ways that are seriously morally wrong or unjust. Such situations are also
exemplars of when anger and distrust are fitting, for someone who acts in a way that is
seriously unjust has performed an action that violates a norm on proper conduct while she
is motivated by ill will or has shown insufficient concern. Further, in many cases in
which the nonconsequentialist views justify harming an offender, the harms that we are
justified in imposing on the offender will be deserved. Since these harms are deserved by
offenders in so many cases, it will be plausible that desert is playing an important role in
the justification of punishment. But, as I have suggested, desert does not even play a role
in the explanation of punishment’s expressive character. What makes the sanctions we
impose on offenders express anger and distrust is that these sanctions are functionally
similar to anger and distrust as well as isomorphic to them. By engaging in these
sanctions, we express our anger and distrust at offenders in a manner that is justified, which is all we can ask of an account of punishment’s expressive character.

I believe the account I defend goes some distance toward blunting the appeal of retributivism as the only theory of the justification of punishment that can respect the fact that punishment is offender-focused and expresses condemnation. While I have not discussed state punishment per se, I believe that with few modifications, the theories I have discussed can be applied in the realm of state action. If I am right, we have understood how someone could consistently hold that justified punishment is offender focused, criminals often deserve punishment, the desert of offenders does not play any role in why we are morally justified in punishing them, and why and when we are justified in punishing people who have been globally manipulated.

5.9 CHAPTER CONCLUSION

One focus of this chapter has been the justifiability of punishing Professor Plum and the neuroscientists. In order to adequately address this issue, we needed to examine the justification of punishment more generally. In an attempt to satisfy that aim, I first turned to the view of punishment as quarantine defended by Shoeman and Pereboom. That view sought to show that it is possible to justify punishment without the moral responsibility of offenders playing a justifying role, but we saw that it was because the quarantine view does not treat moral responsibility as having a justifying role that it does not succeed.

I then turned to Farrell’s theory regarding the justification of deterrent violence and showed that the principle of justice that governs such violence also covers the
imposition of incapacitory social isolation and showed how the view would treat the cases of Plum and the neuroscientists. Finally, we saw that by appealing to the functions of anger and distrust as well as to isomorphisms between their conditions of fit and punishment’s conditions of justification, we could understand how punishment expresses condemnation to offenders.

This chapter shows that it is possible to hold that socially isolating Plum is justified and fitting, but not for retributivist reasons. It thus refutes the final attempt by incompatibilists to use the manipulation argument to threaten compatibilism. I have not argued that we would be justified in punishing Plum, but I have shown the next best thing: we would be justified in socially isolating Plum in a state of nature situation and by doing so we would be expressing fitting condemnation of his action. With little modification, I believe this theory can be applied to punishing Plum in a political society. If so, we can put the manipulation argument to rest.

5.10 DISSERTATION CONCLUSION

It has been quite some time since we first met Pereboom’s Professor Plum case, the most sophisticated global manipulation case on offer. Pereboom uses that case as the basis of an argument that compatibilism is false, but we have seen that his argument does not succeed because the persistent intuitions that Plum does not deserve anger and that he does deserve our sympathy do not show that Plum lacks moral responsibility for killing Ms. White. But I have not just shown that Pereboom’s manipulation case does not support his conclusion; I have shown this by careful attention to our practices of holding each other morally responsible. So my work also helps to inform our understanding of
moral responsibility more generally. I showed that work on the reactive emotions conception of moral responsibility needs to be informed by work on different senses of appropriateness of emotions and that the reactive nonemotions conception could be helpfully extended by showing how trust and distrust are implicated in our moral responsibility practices. Finally, I turned to punishment. While it might seem that the theoretical options for someone who is attracted to a theory of the justification of punishment that is offender-focused boil down to retributivism, I showed that is not true. It is possible to take moral responsibility seriously as a justifying factor in punishment, avoid retributivism, allow that criminals often deserve punishment but that their desert is not a justifying factor in their punishment, and account for the condemnatory aspect of punishment. Even though intuitions about Plum’s case can mislead about his moral responsibility for killing Ms. White, paying close attention to his case has yielded a more nuanced appreciation of moral responsibility and punishment—for that, he deserves our thanks.


