The Southern Journal of Philosophy
Volume 58, Issue 2
June 2020

RESPONSIBILITY: THE STATE OF THE QUESTION
FAULT LINES IN THE FOUNDATIONS

David Shoemaker

Abstract: In this article, I discuss five major fault lines in the foundations of responsibility theory—a relatively new field—fault lines traceable to P. F. Strawson's groundbreaking "Freedom and Resentment." They are about the proper methods and content of responsibility theory, and disputes over these foundational issues have led to a messy and wildly divergent set of theories and approaches in the field. My aim is simply to identify and sort out these fault lines so that we can at least agree on the source of our disagreements, and so that we may also perhaps realize the pressing need to address these fault lines first, in order to help resolve some of our downstream disputes and nurture responsibility theory into a more unified and theoretically sophisticated philosophical enterprise.

Over the last fifty years, theorizing about the foundations of responsibility has become a fractured enterprise, with theorists taking off in multiple and competing directions. What, then, is the State of the Question? The foundations of responsibility are, like America, riven by fault lines.

I will begin with a brief discussion of what I take the *foundational* concerns in our theoretical enterprise to consist in, and then I will say something about their fractured state in the field. I will go on to trace how we got to this point from a single article published in 1962, and I will follow up with

David Shoemaker is Professor in the Department of Philosophy and the Murphy Institute of Political Economy at Tulane University, as well as the director of the Murphy Institute's Center for Ethics and Public Affairs. His current areas of research are agency and responsibility, moral psychology, and the relation between humor and morality. He is the author of two monographs, most recently *Responsibility from the Margins* (Oxford University Press: 2015) and more than sixty papers. He is the general editor of the ongoing OUP series *Oxford Studies in Agency and Responsibility*, an associate editor of the journal *Ethics*, and a cofounder and coeditor (with David Sobel) of the long-running ethics blog *PEA Soup*.

The Southern Journal of Philosophy, Volume 58, Issue 2 (2020), 205–237.
ISSN 0038-4283, online ISSN 2041-6962. DOI: 10.1111/sjp.12370

a more detailed discussion of the five most visible fault lines in contemporary theorizing about responsibility. When I can, I will also gesture toward the possibility of a more unified direction going forward. What will emerge from the discussion is that responsibility theory is still in its early stages, finding its footing, and like the toddler that it is, the field may remain messy and unbalanced for a while. Recognizing, grappling with, and trying to bridge these foundational fault lines will be necessary in order for the field to reach maturity.

WHERE WE ARE

What counts as a foundational issue in a philosophical domain like this? I take the metaphorical language pretty literally: a foundational issue is one on which a theoretical enterprise is built. Most generally, foundational issues have to do with the *basic content(s)* and *method(s)* of the enterprise. Basic content refers to what the theorist theorizes about, that is, the (hopefully) agreed-upon data set. The question of method asks, "Given the basic content of the enterprise, *how* are we to theorize about it? What are the tools we may or may not employ in accounting for the data set?"

If there are fault lines at the foundations of responsibility regarding basic content or method, there is likely to be serious disagreement over everything else. And there is. Of course there is the standard sort of disagreement over which theory best accounts for some data set. But the foundational disagreement I am talking about is ultimately over whether what people are doing even counts as theorizing about *responsibility*.[1] It is disagreement over how we can even *engage* with different theories and theorists, given their different starting points and methods. This disagreement becomes evident when we try to categorize and comparatively assess the theories that have gained some traction in the literature.

To see why, recall the famous book *Four Views on Free Will*, published in 2007. In it, three longstanding theories (and one compelling upstart theory) were laid out clearly and in their most plausible contemporary form, and the authors (John Martin Fischer, Robert Kane, Derk Pereboom, and Manuel Vargas) critically engaged with each other's positions. The views laid out were compatibilism, libertarianism, hard incompatibilism/skepticism, and revisionism. This book was a model introduction to the State of the Question about free will, and it surely taught many students all the basic

[1] This sort of dispute comes out much more explicitly in informal conversations than it does in print. But it has nevertheless become evident in the literature for reasons I will now explain.

information they needed to know in order to situate themselves to do work in the field.[2]

What, then, if we wanted to put together a contemporary companion volume about responsibility? What would be the relevant views we would ask different theorists to write about? Answer: *Who knows?*[3] There simply *is* no agreed-upon way to carve up and categorize the theories we have relative to one another. Here is a partial list of the "theories of responsibility" people have put on the table just over the last sixty years: (a) utilitarian/consequentialist theories (e.g., Smart 1973); (b) agent-causal theories (e.g., Campbell 1957; Chisholm 1964); (c) event-causal theories (e.g., Kane 1996); (d) "ultimate" or "absolute" theories (e.g., Strawson 1994); (e) ledger theories (e.g., Zimmerman 1988; Haji 1998); (f) quality of will theories (e.g., Strawson 1962; Scanlon 1988; Arpaly 2006; Shoemaker 2013, 2015; Arpaly and Schroeder 2014); (g) revisionist theories (e.g., Vargas 2013); (h) conversational theories (e.g., McKenna 2012); (i) expressive/communicative theories (e.g., Watson 2004, 219–59); (j) judgment-sensitive theories (e.g., Scanlon 1998; Smith 2005); (k) reasons-responsive theories (e.g., Fischer and Ravizza 1998; McKenna 2006; Brink and Nelkin 2013; Sartorio 2016); (l) rational abilities theories (e.g., Wolf 1990; Nelkin 2011); (m) fair avoidability theories (e.g., Brink and Nelkin 2013); (n) deep-self or mesh theories (e.g., Frankfurt 1988; Sripada 2016); (o) skeptical theories (e.g., Rosen 2004; Levy 2011; Pereboom 2014); and even (p) *meta*skeptical theories (e.g., Sommers 2012). There are also theories of different individual *types* of responsibility, that is, theories about (q) accountability (e.g., McKenna 2012; Pereboom 2014; Nichols 2015, (r) answerability (e.g., Duff 2007; Smith 2012, 2015), and (s) attributability (e.g., Gorman 2019), as well as (t) theories incorporating all three (e.g., Shoemaker 2015), (u) theories incorporating just two (e.g., Watson 2004; McKenna 2012); (v) theories incorporating and defending only one (e.g., Smith 2012, 2015); and (w) pluralist theories advocating different plural categories altogether (e.g., Mason 2019). There are also theories about the source of responsibility's nature, including (x) response-dependent theories (e.g., Shoemaker 2017), (y) response-independent theories (e.g., Tadros 2005; Brink and Nelkin 2013), and (z)

[2] True, many of the authors of this volume were motivated to develop theories of free will by their desire to articulate the nature of the control condition necessary for responsibility. But theirs were not theories of responsibility itself, and as I will show during the course of this article, there are powerful reasons to think that the right theory of free will will not tell us much of value about responsibility at all.

[3] This is precisely the answer that came to the fore when I began to discuss putting together such a book with Michael McKenna, Dana Nelkin, and Chandra Sripada back in 2016. We simply gave up, in part for some of the reasons detailed in this essay.

metaphysically-interdependent (response-dependent *and* response-independent) theories (e.g., McKenna 2012).

This is alphabet soup, a mess made even messier by the fact that many of the distinctions being tracked are orthogonal to one another and that several of the theories overlap, sometimes heavily (e.g., reasons-responsive, rational abilities, and judgment-sensitive theories). So why is there such a difference, then, between free will's State of the Question and responsibility's?[4]

There is no State of the Question right now in responsibility theory because there is no single—*the*—question. There *was* for free will, namely: "What are the conditions of free will in light of the possible truth of determinism?" This is a longstanding and heavily investigated question, and it was also a focused question, so it generated answers that, while competing, nevertheless agreed on the basic content and method for proceeding. Because responsibility was universally taken to presuppose free will, the assumption was clearly that if we could settle the matter of free will, we could go a long way toward settling the matter of responsibility. This would suggest, then, that the different general theories of responsibility would come to mirror their presupposed theories of free will, being compatibilist, libertarian, skeptical, or revisionist themselves. But that is obviously not what we have on our hands today (although these are among the relevant theories), and that is because the focusing feature of the free will question—about the threat of *determinism*—was removed, for many people, by P. F. Strawson's groundbreaking work in "Freedom and Resentment." Removing that threat resulted in the ushering in of a bunch of new questions, and the mess of theories we see is a result of the attempt by different theorists to answer these different questions in sometimes radically different ways. To see this point, we need to see how we got to this point.

HOW WE GOT HERE[5]

Most pre-Strawson theorists were moved to investigate free will so as to figure out what might morally justify punishment—most frighteningly,

[4] It is worth noting that in light of Frankfurt's 1969 paper "Alternate Possibilities and Moral Responsibility," many free will theorists abandoned their investigation into "the ability to do otherwise" in favor of a control condition on moral responsibility. To the extent that control theorists' investigations are parasitic on their theories of responsibility, they will be similarly riven in the ways I lay out below. (Thanks to Randy Clarke for noting this point.)

[5] Some of what I say in this section was noted (more succinctly!) by Paul Russell (2017, 67–93, esp. 67–68).

God's eternal damnation—and they quickly saw a number of challenges, none more threatening than the bogeyman of determinism, the theory that all events are wholly determined by prior events in combination with the laws of nature.[6] Among those events are human actions. If all of our actions are totally determined, then none of us has free will and it would seem that no punishment or damnation could ever be morally justified. Lack of free will entails lack of moral responsibility. Thus was borne over 2000 years of metaphysical work on free will, resulting roughly in three classical views: (a) free will is still compatible with determinism, insofar as what matters for freedom is just that one's actions are determined by the *right kinds* of causes, and these may include one's own desires and decisions (*compatibilism*); (b) determinism is indeed incompatible with free will, but fortunately determinism is false, so free will is possible (*libertarianism*); or the depressing (c) determinism is incompatible with free will, and determinism is true, so we all lack free will (*hard determinism*).

By the time Strawson came round, compatibilists had mostly come to think of moral responsibility in utilitarian terms: what justifies the punishment of wrongdoers, they thought, are its future-oriented good consequences, for example, its preventing wrongdoers from hurting anyone else for a while and/or its deterrent effects on potential wrongdoers. On this view, determinism is not only no moral threat to our warrant in holding people responsible, it may also be necessary to its success, as what punishment does is instill new causes in people so as to determine their future behavior in more desirable ways.[7]

Of course, viewing punishment in this way is objectionable to many people because it ignores a beloved platitude: criminals must *deserve* their punishment. A utilitarian justification of punishment sends us to jail without desert. Libertarians, on the other hand, have desert in spades. What they lack is a clear or viable metaphysics to back it up, as the conditions for actions to be undetermined—and yet also to be somehow not a function of chance—lead into spookiness quite quickly.

Strawson stepped into this debate and said both sides get some things right, but they also get some things seriously wrong. Utilitarian compatibilists (what Strawson called "optimists") are right that the justification for our responsibility practices is not threatened by determinism (by "the

[6] Initially, God's foreknowledge was the supernatural version of determinism's naturalist threat. The best contemporary challenge to free will along these lines can be found in the Consequence Argument. See Ginet (1966), Wiggins (1973), and Van Inwagen (1983).

[7] For a very clear articulation of this view, see Stace (1953).

facts as we know them"), but they are wrong to think of those justifications in forward-looking, managerial terms. Libertarians ("pessimists") are right about the backward-looking and interpersonal features of responsibility, but they're just wrong that a libertarian metaphysics is needed (or even possible) to justify them.

What Strawson develops is a naturalistic, sentimentalist solution to the problem of responsibility and determinism. He starts by examining our actual noninstitutional, nonpunishing, interpersonal and emotional responsibility-responses (a domain where, he says, the discussion is less crowded with disputants). Let us identify those, he suggests, and then let us identify when we naturally suspend them. He focuses mainly on resentment. If we never suspend resentment for reasons that have the kind of universal human application that would have to be a defining hallmark of determinism, then the theoretical truth (or falsity) of determinism is just irrelevant to our actual responsibility responses and practices.

To explain, my emotional default is to respond with resentment to your stepping on my foot, but if I find out that it was an accident, or that you were pushed or forced, or that you were a "hopeless schizophrenic," a child, or "warped or deranged" (Strawson [1962] 2003, 79), my resentment will naturally be suspended. Now if determinism were true, it would have to apply to—and seemingly *excuse*—everyone always for their actions. But the types of excuses Strawson surveys just *are not* true of everyone always: our actions are not all accidents or coerced, and we do not all suffer from global delusions. Consequently, the types of things that actually—naturally—render our responsibility sentiments inappropriate (or appropriate) have nothing to do with determinism.

Furthermore, there is no external justification for the framework of responsibility-responses and associated practices within which we live. Strawson's basic assumption is that we humans simply find ourselves within this framework of interpersonal life, and it is constituted via our vulnerability to the reactive emotional attitudes (like resentment) associated with responsibility. Without these reactive attitudes, there simply could be no interpersonal life, and, indeed, it is psychologically impossible for us to give them up (and even if it were possible, it would be a terrible idea, given the costs to interpersonal life thereby). Consequently, to ask for a justification for our range of responsibility-responses is just to ask for a justification for being human[8]: it utterly misses the point. The most we can get are justifications *internal* to the framework, justifications rendering specific responses

[8] I owe this phrase to Sean Foran.

felicitous simply in virtue of how our sentimental sensibilities have been built (Darwall 2006, ch. 4). Consequently, and roughly speaking, compatibilism wins, insofar as the truth of determinism is irrelevant to (and so is compatible with) our internally-justified responsibility assignations, but nevertheless (and in contrast with the view of the optimist) this sort of responsibility is thoroughly *backward*-looking, interpersonally engaged, and humanly—sentimentally—infused.

In laying out this approach, Strawson advances several striking new methods and positions that starkly contrast with previous approaches to the issue:

- *The Data Set*: He starts by investigating the natural sentiments at the heart of our interpersonal responsibility practices—our reactive attitudes—whereas previous work focused primarily on institutional punishments, sanctions, and even divine retribution, all of which directly aim to set back people's interests and thus cry out for serious moral justification. Strawson here was also developing a much more empirically-informed approach to the nature of responsibility (Russell 2017, 67).

- *Holding vs. Being Responsible*: According to several interpreters, Strawson advances a view of responsible agency that actually makes it a function of our sentiments, constructed out of the aptness of *holding* people responsible (see, e.g., Bennett 1980; Wallace 1994, 95–109; Watson 2004, 221–27, 2014; Shoemaker 2017). What makes someone a responsible agent, in other words, is their being somehow appropriately regarded as such by others. Previous work presumed that properly holding others responsible is a function of their *being* responsible, that there is an antecedent fact of the matter of whether someone is responsible, and it is determined not by us but by the state of the world independently of us.
- *No Desert, Please*: Strawson argues that the search for an external justification for our responsibility practices has been wrongheaded all along, that there is no such justification (e.g., grounded in rationality, utility, or desert), and that the relevant internal justifications for specific responses are immune to the threat of determinism. Indeed, because our reactive attitudes are merely emotional responses to people's qualities of will, they are also immune to traditional worries about their moral justification or the need for some kind of libertarian-style desert to ground it.

- *Backward-Looking*: Responsible agency is most fundamentally a backward-looking matter, that is, my resentment of you in response to your wrongdoing is made appropriate in virtue of what you did and the attitude you had in doing it, not in virtue of some forward-looking value my resenting you will help to generate or promote. This is in stark contrast to the utilitarian treatments of responsibility popular at the time.[9]

- *Quality of Will(s)*: Our reactive attitudes directly target, and are made appropriate in virtue of, people's quality of will, not their actions as such (which merely *manifest* their quality of will). This is in contrast to previous theories whose focus was directly on people's actions and their ability to *do* otherwise. Given that our responsibility-responses directly track quality of will, then, it is also possible that there are multiple *types* of quality of will which they track, in which case there could conceivably be multiple types of responsibility.[10]

The source of these bullet points is Strawson's guiding view that we have a psychologically-entrenched set of responsibility attitudes and practices that we can appropriately engage in—and investigate—*without worrying anymore about free will or the threat of determinism*. Strawson aimed to free us from the quest for freedom. Many have adopted this view with relief; many others have dug in, rejecting or heavily qualifying it. This dispute has, resultingly, produced five fault lines in the foundations of responsibility corresponding to the five bullet points above, fault lines that have only grown wider in recent years. I turn now to discuss them in more detail.

FAULT LINE 1: THE DATA SET

The one thing that most theorists *do* agree on in Strawson's wake is that, for interpersonal responsibility, we should indeed pay much closer attention to our interpersonal responses, as our responses are tied tightly to, and reveal something important about, responsibility. Indeed, the following biconditional has become something of a platitude amongst (many) theorists:

[9] This is not to deny some forward-looking elements to Strawson's picture. As he notes at the end, "It is far from wrong to emphasize the efficacy of all those practices which express or manifest our moral attitudes, in regulating behavior in ways considered desirable; or to add that when certain of our beliefs about the efficacy of some of these practices turns out to be false, then we may have good reason for dropping or modifying those practices" (Strawson [1962] 2003, 93).

[10] This last point is certainly not explicit in Strawson, but his view opens up this possibility, and I have exploited it in Shoemaker (2013, 2015).

> One is responsible for X if and only if it would be appropriate to hold one responsible for X.[11]

The obvious first question to answer in figuring out how best to understand this "platitude" is what the relevant responses are, that is, what does holding someone responsible amount to? What we are standardly told here is that the negative or positive responses for which responsible agents are eligible are *blame and praise*, and that these are constituted by the Strawsonian reactive attitudes. Strawson himself lists several, including "such things as gratitude, resentment, forgiveness, love, and hurt feelings" (Strawson [1962] 2003, 75). But he goes on to focus exclusively on resentment, and then (by analogy) on indignation and guilt, what I label the Holy Trinity of blaming emotions.

There are at least three major disputes that have arisen in recent years about the nature of this data set, however, and these disputes have led to what I believe is a methodological morass in the field. Perhaps I am an alarmist about this point in part because I have been the genesis of some of these disputes myself. But I honestly believe we cannot move forward as theorists of responsibility without facing up to these serious challenges. The three disputes generating the first general fault line about our data set are over (a) what emotions are to be included in the set of responsibility-responses, (b) what role emotions are to play in holding people responsible, and (c) whether emotions are necessary at *all* to our responsibility-responses. I briefly discuss each in turn.

A. Beyond Resentment?

Almost everyone theorizing about our responsibility-responses starts with the reactive attitude Strawson did, namely, resentment, and this is why Quality of Will theories of responsibility have become so popular in the literature.[12] But what if Strawson had focused on *hurt feelings*, another reactive attitude he mentions, instead of resentment? As I have recently argued, had he done so we would not have anything like a Quality of Will theory of responsibility on our hands (Shoemaker 2019). That is because *pure* hurt feelings (*sans* resentment) are very often not an appraisal of the hurter's

[11] This formulation is drawn from Wallace (1994, 91). Something like it has been adopted by many since, including Fischer and Ravizza (1998, 7), Arpaly (2006, ch. 1), McKenna (2012, ch. 2), Brink and Nelkin (2013, 287), and me, in Shoemaker (2017).

[12] See, e.g., Scanlon (1988), Arpaly (2003, 2006), Smith (2005), Harman (2011), McKenna (2012), Talbert (2012a, 2013), Shoemaker (2013, 2015), Arpaly and Schroeder (2014), Hieronymi (2014), Sripada (2016), and Bjornsson (2017).

quality of will, at least in Strawson's sense. Hurt feelings instead are aptly generated by what others *think or feel* about us, where what they think or feel is worse than we had previously thought. This means someone may think or feel something about us in a way that legitimately hurts our feelings, even though the thought or feeling is accidental, coerced, or completely justified, and even when the hurter has the highest goodwill, affection, and esteem for us. That is, hurt feelings do not tend to be suspended for *any* of the same reasons Strawson surveyed as naturally suspending resentment. This is most obvious in cases of telling hard truths in close relationships. Sometimes hurtful things must be said. For example: "Your drinking is out of control," or "I've fallen in love with someone else." Hurt feelings also arise in relationships where one person's understanding of the terms of the relationship is just different than the other's. If you—someone I think of as a mere friendly acquaintance—give me a gift I have no use or room for, and I justifiably throw it away, you—someone who thinks of me as a good friend—may nevertheless be legitimately hurt when you see your gift in my trash.

Despite their occasional blame-like feel, though, hurt feelings are not blame. They also are not praise. Indeed, it is unclear just how to characterize them. Nevertheless, they very much seem to be *responsibility* responses. After all, they attribute some attitude to the hurter's *agency*. It is not as if your nonagential features—your height, weight, or eye color—can legitimately hurt my feelings. Rather, there must be something about you *qua* agent that does so, and that seems to put hurtful attitudes into the responsibility hopper alongside other agential activities and attitudes for which we are more familiarly responsible. In addition, hurt feelings seem to call for a kind of response from hurters that also seems quite familiar in the responsibility-domain, namely, acknowledgment, apology, guilt, remorse, and/or recompense (Shoemaker 2019, 144). If I know that I will hurt you when justifiably telling you a hard truth, I may well need to apologize as I do so.

The problem caused by hurt feelings stems from this dilemma: either they do or they do not count as part of our data set of ways to hold people responsible. If they do, then because they do not necessarily implicate Strawsonian quality of will, all those theories saying that responsible agency is all and only about one's quality of will are false. But as it turns out, the basic capacities required by virtually every other theory of responsibility are also unnecessary. That is because all that is necessary to be a hurtful agent is that one has the capacities to have thoughts and feelings about others, *and that is it.* This means that such responsible agents need to meet neither the standard control condition (sometimes called the "free will" condition)

nor the epistemic condition nearly universal in theories of responsibility. I can hurt your feelings—and you can aptly respond by being hurt—without my necessarily being able to do otherwise in any sense, and without my necessarily knowing that what I am doing, saying, or thinking is hurtful.

On the other hand, if we exclude hurt feelings from our data set, we are likely to be doing so precisely because they *do not* require agential control, knowledge, or quality of will. But then we are categorizing our reactions as responsibility-responses or not in virtue of antecedent conceptual considerations about what counts as *responsibility*. But on what basis might we do so? After all, there are serious disputes about this very concept. To make your case for one construal over the other, it is hard to see how you could proceed except by appeal to specific cases and, in particular, to what your and our reactions are in those cases. But then you are leaning on our responses to determine the contours of your concept. So, we cannot rule out hurt feelings without begging the question. But if we include them, we cause serious disruption to the search for a unified concept or theory of responsibility.

B. Beyond Human?[13]

What role are the emotions supposed to play in our responsibility-responses? On Strawson's view, they are central to our understanding of responsibility, and he even claimed, famously, that "only by attending to this range of attitudes can we recover from the facts as we know them a sense of what we mean, i.e., of *all* we mean, when, speaking the language of morals, we speak of desert, responsibility, guilt, condemnation, and justice" (Strawson [1962] 2003, 91). But how crucial are these emotional responses *really*? Are they contingent or constitutive? Do they merely cast light on our understanding of responsibility, or are they essential to it in a way that renders responsibility anthropocentric?

One way to think about the question is to consider various thought experiments that move us further away from the responsibility system as we currently know and experience it. Do we need a community of emoters, of people who emotionally hold one another responsible, in order for there to be responsible agency (see McKenna 2012, 107–8; see also Russell 2017, 46–96)? Derk Pereboom (2001, xx–xxi) imagines a race of beings who are rational but emotionless, and who care about right, wrong, and holding each other to account, but they do so without resentment, indignation, and guilt. Is their conceivability not enough to establish the possibility of responsible

[13] Thanks to Michael McKenna for his contributions to this section.

agency without emotionality? Why is the motivational basis provided by the emotions that we humans may need to hold others to account actually necessary to responsible agency, full stop? (McKenna 2012, 110–13).

These questions are good and difficult. They go to the heart of whether holding responsible and the nature of responsibility are distinctively human, or if they are somehow independent of humanity, perhaps instead being a function of mere rationality, so that gods and angels could be responsible as well. Does our data set for theorizing about responsibility stop at the limits of our humanity? Obviously answering the question in one way or another will yield very different theories of responsibility.

C. Beyond Blame and Praise?

Most people consider "holding responsible" in the biconditional to refer exclusively to blaming or praising responses. Indeed, some have construed the Strawsonian enterprise as establishing responsible agency just in case some blaming or praising reactive attitudes to the agent are apt.[14] But people actually seem to be responsible for some activities *without* blame or praise being implicated in any way. Holding responsible in such cases typically involves confrontation, and sometimes mere complaints or worries. We are indeed holding someone to a demand, expectation, or hope in these cases. But our responses do not have a blaming (or praising) cast to them.

Gary Watson gives several compelling examples. "You complain to your partner that she hasn't been giving you enough of her time or doing her share of the housework" (Watson 2019, 224). This is, of course, an accusation of wrongdoing, but responsibility complaints need not be restricted to wrongdoing, for example, "'Do you have to play the radio so loud? I can't work.'" (Watson 2019, 224). And there need even be no accusatory element to such demands: "You seemed so distant last night. What's up?" (Watson 2019, 224). And we should not forget Watson's early famous examples of Gandhi and King, who vigorously held their oppressors to account without blame (Watson 2004, 257–58, 2019, 225). These are demands for acknowledgment, surely, but nothing more (Watson 2019, 225; see also Shoemaker 2015, ch. 3).

[14] This is how Scanlon interprets Strawson, for one: "Strawson does not describe reactive attitudes as forms of blame, but this identification is a natural application of his analysis" (Scanlon 2008, 224n6). How does one account for responsibility for morally neutral actions, though (e.g., taking a shower in the morning)? One goes subjunctive: *were* the action morally loaded, one would be an apt candidate for blaming or praising reactive attitudes.

Now what do these examples of holding-responsible-without-blame-or-praise mean for our theorizing? First, we might deny that these count as responsibility responses (and so deny that they give us *any* insight into responsible agency). Some might lean towards this option either because these examples are not *moral*, or because some of them are not examples of *wrongdoing*. Others might lean toward this option because they believe that responsibility—the very concept—must implicate blame or praise. Both leanings seem ultimately to beg the very question at issue, though.

The second option one might take is to accept that these confrontational complaints, protests, and encounters do count as holding-responsible responses. But doing so once again opens the door to a fractured concept of responsibility. Being responsible would thus *sometimes* render you eligible for blame or praise, but sometimes it would not. We would thus again have to give up the biconditional that most theorists love.

* * * * *

The three disputes here have been over the status of the biconditional, specifically about the nature of our data set in holding one another responsible: What counts as the relevant way of "holding responsible" and whose responses count? Indeed, what are we supposed to be theorizing about in the first place? The data set we use will obviously shape the theory of responsibility at which we arrive. But then how can we choose or filter the data set except on the basis of an antecedent concept or theory of responsibility?

This first fault line stems from disputes over the nature of the data set and how exactly to determine its contents. The second stems from disputes about the data set's role.

FAULT LINE 2: BEING AND HOLDING RESPONSIBLE

Let us suppose, magically, that we can establish agreement on what the relevant responsibility-response data set is. We come now to our second fault line, namely, over what role those responses are supposed to play in our theorizing about responsibility. There are two competing methods here. On the one hand, we can take our data set to constitute an *epistemic aid*. Our responses are pretty good (but not perfect) trackers of the facts about responsibility,[15] we might say. On this approach, deployments of the

[15] Facts which could, but need not, be facts about free will.

holding-responsible data set are subject to revision in light of the correct theory of responsibility, whatever that turns out to be. So, for example, if what we believe is the true theory of responsibility requires capacities for normative knowledge that those with certain cognitive disabilities do not have, then any responsibility-responses to them will be rendered inappropriate by the facts of responsible agency, regardless of whether many people *do* continue to respond to the cognitively disabled in a responsibility fashion (see Fischer and Ravizza 1993, 18).

On the other hand, we might appeal to our responsibility-responses as they stand in order to *construct* our theory of responsibility. The source of this method is once again Strawson. As Gary Watson writes:

> Whereas traditional views have taken these [reactive] attitudes to be secondary to seeing others as responsible, to be practical corollaries or emotional side effects of some independently comprehensible belief in responsibility, Strawson's radical claim is that these "reactive attitudes" . . . are *constitutive* of moral responsibility; to regard oneself or another as responsible just is the proneness to react to them in these kinds of ways under certain conditions. (Watson 2004, 220)

The fault line here, which is growing steadily in the literature, is over *response-dependence vs. response-independence* about the nature of responsible agency. Return to the biconditional: One is responsible for X if and only if it would be appropriate to hold one responsible for X. The dispute here is over which side of the biconditional (if either) is more fundamental than (that is, has metaphysical priority over) the other.[16] Response-independent theorists say it is the left side: it would be appropriate to hold one responsible for X if and only if *and in virtue of the fact that* one is responsible for X. Response-dependent theorists reverse the priority: one is responsible for X, they say, if and only if *and in virtue of the fact that* it would be appropriate to hold one responsible for X (see Todd 2016; Shoemaker 2017, forthcoming).

It is hard for some people to make sense of response-dependence about responsibility. Indeed, how can we make sense of holding others *responsible* unless they already *are* responsible? Does our holding them responsible not in and of itself make their being responsible more fundamental than—and fix the appropriateness of—our holding them responsible?

It is not so hard to understand response-dependence, though, at least in other value domains. Take amusement, for example.[17] It is likely that,

[16] An important alternative I lack space to discuss here is that neither grounds the other; see McKenna (2012, 50–55).

[17] This is an analogy discussed by Todd (2016) and Shoemaker (2017).

because of how we have been built, what makes something funny is in some way a function simply of what we find amusing. This does not seem so bizarre, given the wide and seriously disparate variety of things we find funny (e.g., slapstick, puns, satire, ridicule, etc.). What could unify them all except that they are just the sorts of things we tend to be amused by? But then there is nothing about amusement in and of itself that presupposes some antecedent concept or account of the funny. Of course, it is true that in being amused I find something *to be funny*, but we could simply describe this attribution as a conceptual overlay onto the more fundamental notion of amusement. In other words, we could describe "the funny" as just referring to *whatever people are amused by*, and "finding funny" we could simply describe as shorthand for "being amused by."

Of course, the purely dispositional response-dependent understanding of the funny as "whatever people are amused by" cannot be right as it stands, for it lacks *normativity*. It is not that when someone is amused we think that what she is amused by is funny in virtue of her actual amusement. After all, she may be amused at the unamusing, as in when she is stoned, when she is giddy with exhaustion, or when she is in a stressful situation like a funeral (D'Arms and Jacobson 2014). And even when these mind-altering conditions are not in place, people may be *incorrectly* amused by bad jokes. Some people, after all, just have bad senses of humor.

Consequently, a more plausible response-dependent account of the amusing must say that what makes something funny is ultimately a function of what people *appropriately* find amusing. Of course, our account of what is appropriate cannot rely on response-independent facts about the funny! Instead, the picture will have to be something like the following: Amusement is appropriate in response to properties that refined human humor sensibilities have been built to respond to with amusement (Shoemaker 2017, 487–90). In other words, what makes certain objective properties the funny-*makers*—the properties to which it is appropriate for us to respond with amusement—is our properly functioning sense of humor.[18]

By analogy, then, it may be that, because of how we have been built, what makes someone a responsible agent ultimately depends on our properly functioning *sense of responsibility*, which is what issues in our reactive

[18] There are, predictably, complications here, most importantly over what a refined or properly functioning sensibility is. See Shoemaker, forthcoming, for ways of addressing that issue and other complications. Note also that, if one goes in this direction, one will likely think holding responsible, and so responsibility itself, is a distinctively human enterprise, maintaining relativism about responsibility at the species level.

attitudes. Suppose, for example, that a kind of anger is the paradigmatic emotional core of blame. We could thus appeal to anger to construct a response-dependent version of responsibility as follows:

> What makes someone blameworthy, and thus responsible, for some action or attitude is just that she is the appropriate target of blaming anger. Blaming anger is appropriate in response to *slights*. But one cannot understand what slights are without essential reference to our anger sensibilities; that is, slights are (all and only) the agential activities to which our refined anger sensibilities have been built to respond with blaming anger. (Drawn from Shoemaker 2017, 508–12)

A full-fledged response-dependent theory of responsibility would of course have to show how this structure would be true, not only of anger, but of the remaining reactive attitudes, which is no mean feat. But this might at least be a start.

As noted earlier, response-independence seems by far the more natural stance on the priority question. But there are compelling arguments in favor of response-dependence, starting with the thoroughly natural and deep-seated commitments associated with our reactive emotions (Strawson's point), as well as the extreme difficulty (if not impossibility) of establishing a unified and truly response-independent understanding of the various objective properties in which responsibility ostensibly consists, including *control*, *knowledge*, *voluntariness*, and *quality of will* (Shoemaker 2017, 498–508).

I will not go into these arguments here. Rather, I want simply to note why it matters so much that there is a genuine fault line on this point. It is a methodological dispute, a disagreement over what role the data set—our collection of responsibility-responses—ought to play in our theorizing. On the one—response-independent—hand, those responses are nothing but an epistemic aid to the objective truth about responsibility; on the other—response-dependent—hand, they are the very materials out of which responsibility is constructed. If response-dependence is true, we thus need to start our theoretical investigations by surveying our various emotional responses and figuring out their appropriateness conditions. This will require some down and dirty empirical work, as we are trying to understand our actual human emotions, as well as the sorts of reasons we find compelling to have or suspend them. If response-independence is true, alternatively, we may keep our empirical hands clean, perhaps simply by engaging in the kind of a priori conceptual reflection about responsibility and its nature that has kept philosophers busy since Socrates. This split

between more and less empirically informed philosophy, of course, is itself reflective of a larger split in the discipline.[19]

FAULT LINE 3: "DID YOU SAVE ROOM FOR DESERT?"

Despite the fault lines over the nature of the data set, it is beyond any doubt that at least the most *interesting* forms of responsible agency are about activities that render agents vulnerable to praise or blame. Why interesting? Because of the longstanding concerns that spurred investigations into free will in the first place. Blame, for instance, can *harm*, and harm requires moral justification.[20]

There are many disputes here that contribute to a general foundational fault line. The first dispute is over the nature of blame itself, which actually breaks down into two subdisputes. First, people seriously disagree over whether blame is best construed in *contentual* or *functional* terms. That is, are we to understand blame most fundamentally as consisting in some sort of attitudinal or behavioral content (like anger or relationship-modification), or are we to understand it as consisting most fundamentally in a function, so that certain attitudes or behaviors only count as blame when they play that functional role?[21] Of course, blame can harm regardless of its fundamental constitution, whether content or function. But harm may also not be necessary in either construal of blame's nature, so the determination of whether one has to deal with the moral burden of justification is a result only of which side one takes in the *second* subdispute, namely, about which specific content or function best characterizes blame's nature.

In admitting that not all blame—whatever it is—is harmful, some theorists, in order to keep their eyes on the most interesting prize, narrow their focus to instances of blame that *are* harmful, or at least risk harm, to their targets, such as resentment expressed directly to an offender. This is what Michael McKenna calls "directed blame" (McKenna 2013; see also

[19] I am grateful to Doug Portmore for discussion on this point.

[20] Moral justification may also be needed for praise, even though it typically does not harm anyone, so justifications here tend to be about possible unfairness to the overlooked praiseworthy people when someone undeserving is praised instead. See Watson (2004, 283–85); and Nelkin (2008). I will focus in the text just on blame, as is customary.

[21] On the blame-as-content side of the map are people like Glover (1970), Haji (1998), Zimmerman (1988), Wallace (2011), Scanlon (2008), Wolf (2011), Arpaly (2006), Hieronymi (2004), Sher (2006), McKenna (2012), Darwall (2006), and Macnamara (2013). On the blame-as-function side of the map are people like Smith (2013), McGeer (2013), and Shoemaker and Vargas (2019).

Pereboom 2014; Rosen 2015, 67–68, for blame's being "harsh").[22] What basis could there be to morally justify such an activity?

The fault line here is over hurtful blame's moral ground, and there are three general positions on it. The first has been the most popular, positing that blame's moral justification rests on the *basic desert* of its targets (Feinberg 1970; Strawson 1994; Pereboom 2001, 2014; Clarke 2005; Fischer 2007, 82; McKenna 2012; Scanlon 2013). On this view, those who have knowingly and willingly done wrong, say, are blameworthy, and to be blameworthy is to deserve blame in some basic sense, solely in virtue of having knowingly and willingly done wrong. There exists no further or deeper justification.

The second position is that there *is* a deeper justification, and it is found in a moral theory like consequentialism or contractualism. Perhaps, to draw from the latter, one deserves blame for wrongdoing as a function of principles with which one would have agreed to conform (see, e.g., Rawls 1971, 103; Smart 1973; Lenman 2006; for helpful discussion of "non-basic desert," see McKenna 2012, 161–64).

The third position denies the need for a *moral* basis at all, and so denies the need for desert in responsibility, even in the directed blame cases where someone gets hurt. But how can no moral justification be needed when harm is involved? One possibility is that blame is a kind of appraisal-response to various norm violations that is governed by reasons of an entirely different type than those of morality (that is, reasons that track some kind of *non*moral value). For example, reasons governing blame emotions may simply be reasons pertaining to the accuracy of their appraisals, labeled reasons of *fit* (e.g., D'Arms and Jacobson 2000). On this construal, blame emotions are fitting just in case their appraisal of blameworthiness is "correct" (D'Arms and Jacobson 2000; Shoemaker 2015) or their constitutive thoughts are "true" (Rosen 2015). Desert need not have anything to do with it (see Scanlon 1998, 274–77; Hieronymi 2004, 2019; Smith 2019).[23]

[22] There is a fascinating recent movement afoot to argue that self-directed blame is more fundamental to our understanding of blame than other-directed blame (led by Carlsson 2017, 2019). This is because, while other-blame can be kept private, and so need not hurt anyone, the self-blame of guilt cannot be kept private, and so to the extent a moral justification is needed for blame, it is needed most fundamentally in the self-directed case. Because blaming others aims for them to feel guilt as well, other-blame's justification is parasitic on self-blame's. As McKenna rightly notes (private correspondence), even if true, this view does not move us off of the main question, which is how those cases of directed other-blame that *do* hurt are to be justified.

[23] Alternatively, perhaps basic desert is just nothing more than a matter of fittingness itself. In that case as well, talk of desert may be superfluous and there is no problem of moral justification to be solved. See Shoemaker (2015, 220–23).

Of course, if you believe that one of blame's constitutive thoughts is about how the offender *deserves to suffer* (even if just the pain of guilt),[24] then even on a fittingness construal of blame's appropriateness, desert might still have to play a morally justifying role (Rosen 2015, 82–83).[25] But why should we think that this is the proper construal of blame, "that in paradigmatic cases of resentment, some sort of pain is wanted" (Rosen 2015, 82), where angry blamers must also be thinking that the offender *deserves* such pain? Angry blame surely wants *confrontation*, I admit; it wants to make its anger known to the offender. But to what end? What successfully resolves it? On what I think is the most plausible construal, it is when the offender painfully acknowledges what he did to you, acknowledges how he made you feel in offending you (Rosen 2015, 82; Shoemaker 2015, ch. 3; Fricker 2016). But then why think that the pain is blame's *aim*, the thing that blame wants? Suppose you think the Holocaust was no big deal, so I demand that you go see a powerful Holocaust documentary. My aim in so doing is explicitly to get you to acknowledge its horrors. Suppose that you do. An essential part of your horrified acknowledgment will be its associated pain, not only from your sympathy with those millions who were murdered, but also from your remorse at having overlooked or dismissed those horrors before. Now in having demanded your acknowledgment, have I wanted, or aimed at, your *pain*? Of course not. True, I know full well that if my demand is to be successfully met, you will in fact feel pain, but that is merely an unintended side effect of your acknowledgment. It is akin to my demanding that you go back to the store to pick up the milk you forgot, where an unintended side effect of your doing so is that you will wear down the car's tires a bit. The wearing down of the tires needs no real justification, or if it does, it is established easily by the proportionally much more serious reason for getting milk (the baby needs it, perhaps). So, too, your painful acknowledgment of the horrors of the holocaust itself may need no moral justification either, or, if it does, its justification is easily provided by the proportionally much more serious reasons for generating your acknowledgment. Why, then, cannot the same be true for angry blame, which also demands acknowledgment-that-is-painful?

[24] Many these days hold a view somewhere in this neighborhood. See, e.g., Carlsson (2017, 2019), Clarke (2013, 2016), Duggan (2018), McKenna (2012), Portmore (2019), and Rosen (2015).

[25] Clarke and Rawling (Manuscript) have recently proposed a slightly different view, according to which the conditions that render blame fitting also render it deserved. On this view, even if fittingness itself is not a moral relation, blame is fitting if and only if it has the moral justification provided by desert.

None of this to deny that some people, when blaming others, *do* aim directly at their pain. People can be fairly mean in their retribution. But that is not what angry blame itself wants; it is instead what angry *people* sometimes want. The fact that some people misuse blame for immoral ends should not be surprising. Praise can be used to manipulate people to like and do things for you too, but that does not mean we need a moral basis for praise; it only means a moral justification is needed for *aimed-at manipulation*. So, too, all we need (on the skeptical approach I have sketched) is a moral justification for *aimed-at* pain, not angry blame or pain as such.

My heart obviously lies in this last approach. But I admit that this is a massive fault line, as most others' hearts remain on the desert side of the gap. The reason this is such a big foundational disagreement is that it is over the role, if any, to be played by metaphysical pursuits in our investigations into responsibility. It is the thought that we need an essential appeal to desert (or at least basic desert) for responsibility to be morally justified that has motivated investigations into the nature of free will lo these many years. If we think apt blame has to be deserved, and it is natural to construe desert as requiring agential freedom or control, then off we will go to the metaphysical races. But if apt blame needs no *moral* justification (or needs only a tiny bit), or perhaps if it needs no desert (or if desert simply reduces to fittingness), such metaphysical thickets may be avoided by going with a purely normative approach to blame and responsibility, an approach which sticks with and expands the Strawsonian agenda by identifying and critically examining the internal "normative felicity conditions" (Darwall 2006, 4–5) of our extant responsibility responses and practices, responses and practices which are orthogonal, and likely immune, to worries about the metaphysics of freedom (see also Mason 2019, 7).

Of course, there is no lack of dispute even between those who *do* agree that we need to engage in the metaphysics of freedom, but the disputants here are the familiar compatibilists and incompatibilists. There are also familiar disputes about what the relevant sort of freedom or control consists in, whether it be a mesh between one's deep and superficial selves (e.g., Frankfurt 1988, 11–25; Watson 2004, 13–32; Sripada 2016; Gorman 2019), a responsiveness to reasons (see, e.g., Wallace 1994; Fischer and Ravizza 1998; McKenna 2012; Vargas 2013; Sartorio 2016), or the ability to do otherwise (for the most recent version of this view, see List 2019, 23–24, 97–103). However, these more familiar disputes are not strictly about the foundations of responsibility, I think, for they at least seem to agree on a general method and theoretical content: desert, control, and so free will are fundamental to the enterprise, playing an essential role in the moral

justification of blame's harm, and so we need to start our investigations on the conceptual side of the map, getting clear on the nature of agential freedom and control before doing the metaphysical work needed to see if these abilities are compatible with determinism. Those on the other side of this fault line reject virtually every clause in the previous sentence—that blame is necessarily harmful, that desert and metaphysics are necessary to understanding responsibility—and so they deploy a fundamentally different methodology, starting with at least partially empirical investigation into our actual responsibility responses and practices in order to explore its normative side, that is, the nature and adequacy of the reasons we offer to one another for or against various responses within these practices (again, see Mason 2019, 7). These are indeed radically different approaches, and it is unclear how they may ultimately be brought together.

FAULT LINE 4: BACKWARD-LOOKING OR FORWARD-LOOKING?

One of the positions Strawson was at pains to radically scale back or undermine was the exclusively forward-looking view of "responsibility," according to which our punishment practices, say, are justified in terms of their good effects, both in improving the character of the punished and in deterring other tempted parties. The standard objection to this sort of utilitarian management system is that it is unjust insofar as it has no in-principle objection to punishing the innocent if those same good effects can also be established thereby. Strawson's objection is different, though, namely that this sort of "objective" stance toward our fellows is too impersonal: in treating people like objects to be managed or handled in this way, we are withdrawing from them the kinds of engaged emotional responses that are constitutive of genuine interpersonal life. Forward-looking justifications for responsibility thus rob us of our interpersonal *humanity* (see, e.g., Strawson [1962] 2003, 74, 90, 92–93).

This is a fairly devastating criticism. But it applies only to a flat-footed first-order utilitarian treatment of responsibility. In recent years, though, much more sophisticated utilitarian treatments have made their way into the theoretical mix (whose source may be in Railton 1984), generating the wedge of another fault line. These are two-level theories, akin to Rule Utilitarianism, and they allow that, while our engagement with one another at the first-order can be fully backward-looking, so that our blame and praise of one another is justified or apt wholly in light of what the blamed or praised person *did* (i.e., they *deserve* such a response or it is *fitting*), the

general *practice* itself is justified on forward-looking grounds for the good consequences that having such a system will generate. Manuel Vargas, for instance, argues that our responsibility-system may be justified in terms of its efficacy in "building better beings," agents who are more reasons-sensitive than we are, for instance. But this second-order justification may license (or require) revision of some aspects of our first-order practices, to the extent that they do or do not contribute well already to the second-order aim of the practices (Vargas 2013). And Victoria McGeer makes the more radical case that Strawson *himself* was a consequentialist, that "he thinks of the reactive attitudes as serving a *forward-looking regulative purpose*. . . " (McGeer 2014, 73). And as a matter of fact, Strawson explicitly allows revisions of our responsibility practices to the extent that they are not efficacious in "ways considered desirable" (Strawson [1962] 2003, 93).

On these forward-looking views,[26] there is an external justification for the practices themselves, but it is not at the level that worried Strawson, the level in which a justification for a specific instance of resentment would have it that it regulates the resented agent's behavior better in the future. That sort of treatment *would be* inhuman.[27] A sophisticated consequentialist approach, by contrast, can allow for our first-order responses to one another to be emotionally and interpersonally engaged, and thus backward-looking, as long as that *system* serves some future-oriented good. It is thus not clear why backward-looking and forward-looking theorists must conflict; after all, they could be just talking about different levels, and there is certainly nothing incompatible between the views as they stand. It might thus be thought doubtful that there is any real fault line here.

The fault line comes out, though, when we reintroduce the *first* familiar worry I mentioned above for utilitarian justifications for punishment, which was that they contain no in-principle constraint against punishing the innocent. This worry, now applied to the sophisticated consequentialist, is that there is nothing in principle constraining us from revising our practices so as to allow "apt" blame for the innocent, just as long as such a system generates whatever good consequences or desirability conditions we think ground them. Of course, it may well be that no *actual* system allowing blame of the innocent would ever in fact be justified, given what we take to be desirable. But that is not the objection, which is instead about the fact that the sophisticated forward-looking view does not rule it out *in principle*, and

[26] Others amenable to the forward-looking project are Fricker (2016) and Tsai (2017), both of whom draw heavily from Williams (1995, 35–45).

[27] It is also a seriously "wrong kind of reason" (D'Arms and Jacobson 2000).

it is the in-principle rejection of blaming the innocent that unites many backward-looking theorists. One has to have *done* the thing one is being blamed for, first and foremost, in order for such blame ever even possibly to be appropriate. It cannot be inappropriate as a merely contingent fact.

The problem arises because forward-looking views virtually all deploy Bernard Williams's notion of "proleptic blame," which is a kind of blame for people who lack, in some sense, the reasons there are for them to do other than they do. Blaming these people ostensibly highlights to them the reasons they cannot see—albeit reasons they are *capable* of coming to see in the future—so the "blame" itself is justified in virtue of its (first-order) regulative function. Now because they were blind to these sorts of reasons when they acted, strictly speaking they are not (backward-looking) blameworthy for what they did as they did not meet the standard epistemic condition for responsibility. But they may be (forward-looking) "blame"-worthy, insofar as the blame will at least make salient to them in the future the reasons to which they were blind in the past (Williams 1995, 40–44).

Nevertheless, proleptic blame starts us disturbingly down the path toward licensing blame for the innocent, as there is no reason to think that the innocent will not also respond well to such proleptic mechanisms, by our introducing in them via blame certain reasons to which they *too* had been blind (even though that blindness was irrelevant to what they blamelessly did). This is a seriously slippery slope.

Perhaps the sophisticated theorist could nevertheless justify an absolute prohibition on blame of the innocent insofar as any system of blame that had an absolute ban would be more desirable than an otherwise identical system with only a contingent ban. But now the sophisticated view is subject to the charge of rule-worship that fells most versions of rule consequentialism, namely, it has us putting in place an absolute rule, where the justification of good consequences at the level of the system demands that all first-order rules be ultimately contingent, subject to variation in light of the potential consequences of widespread adherence to or internalization of the rules. You cannot have it both ways.

There are many more things that could be said on behalf of forward-looking theorists, but I simply wanted to say enough to reveal the fault line that stands between them and backward-looking theorists. Again, it should be obvious where my own heart lies, but the issue remains contentious.

FAULT LINE 5: MONISM OR PLURALISM?

Perhaps the most familiar foundational fault line these days is over whether or not there is more than one type of responsibility. As I noted earlier, this

fault line is generated by Strawson's focus on quality of will as the agential feature in virtue of which various reactive attitudes are appropriate or not. "Quality of will" is multiply ambiguous, however, and as different explications of "quality of will" reference different agential capacities, pluralism about responsible agency may beckon (Shoemaker 2015, Introduction).

The original pluralistic line stems from Gary Watson, in his seminal paper, "Two Faces of Responsibility" (Watson 2004, 260–88). Watson was concerned to rebut one part of Susan Wolf's multi-pronged attack on so-called "real self" views of responsibility (Wolf 1990, ch. 2). According to such views, we are responsible for all and only those actions manifesting our deepest selves, those features of our psyche taken to be privileged—because authoritative—in expressing who we really are. For some, these are our higher-order desires, for others our cares, and for others our evaluations of what is best.[28] Wolf's worry was that, on this construal, criticizing me for some bad action is, in its way, like criticizing a car's bent axle for the noise the car makes: the criticism attributes to me a merely causal role in the explanation of some event, and so is a response to a merely superficial form of responsibility. Deep—real—responsibility, on the other hand, involves my being criticizable "on the basis of the role that [I play]," where the blame goes to my moral qualities "in some more focused, noninstrumental, and seemingly more serious way" (Wolf 1990, 41).

Watson showed in response, however, that we actually take up two perspectives on responsibility, perspectives which track different agential features and reveal that we take there to be two "faces" of the concept.[29] First, we have responsibility responses targeting an agent's self-governance, that is, responses to manifestations of an agent's evaluative ends. Our responses here are *aretaic*, appraising psychic features we typically describe as character traits, for example, admiration in response to generosity, kindness, and sympathy;

[28] For the explicit targets of her criticism, see Frankfurt (1988, 11–25), and Watson (2004, 13–32). For more recent expositors and defenders of some form of real self views, see Sripada (2016), Shoemaker (2015, Ch. 1), and Gorman (2019).

[29] "Faces" is notoriously vague. Are these different conceptions of responsibility? Different types? Different notions? Different sides of the same coin? Different features of one thing as detected from different perspectives? Remarkably, in his very recent reflections on this essay in "Second Thoughts," Watson apologizes for using the term, "which is admittedly elusive, and has proven misleading" (Watson 2019, 217). What he was grasping for, he says, was a way of distinguishing between various types of responses and appraisals we might have to someone's attitudes or conduct, all of which fall under the rubric of our "responsibility-responses." Different subsets of those responses generate different normative commitments and different philosophical controversies. These differences reflect different perspectives and motivations we may have when thinking about agential responsibility. For Watson, full-fledged responsibility includes both faces, but it has a "double aspect" (Smith 2015, 121; see also Watson 2019, 217n3).

disdain in response to cruelty and cowardice.[30] Second, there are "responsibility *to us*" responses, responses *holding* agents responsible for their failures in what they owe to us. This second face presumes that the conditions of the first face have been met, but what it adds is that the agent violated some sort of legitimate demand, in a way that was somehow avoidable. For blame of this second sort of be fair, it is often thought that the agent must have had some *control* over the blamable activity (Watson 2004, 264–80).

The first face is the *attributability* (or self-disclosure) face of responsibility; the second is its *accountability* face. The first more or less captures the features emphasized by the real self view. But it avoids Wolf's criticism by nevertheless being a deep form of responsibility. After all, if one's practical identity is a function of one's evaluative commitments, then of course manifestations of that evaluative stance are manifestations of oneself in the world, and appraisals of those expressions are appraisals of oneself. It is hard to see how such responsibility appraisals could be much deeper. As Watson puts it, "Because aretaic appraisals implicate one's practical identity, they have ethical depth in an obvious sense" (Watson 2004, 271).

On Watson's view, responsibility has two general aspects, and our set of responses includes different subsets that track those different aspects, as they reflect distinct ethical interests, namely, those involved in living a good life (ethics) and those involved in behavioral regulation and justice (morality) (Watson 2004, 286). But accountability, for Watson, includes attributability as one of its necessary conditions, so I actually think it is misleading to say that his was a pluralistic view. What I have tried to show in my own work (Shoemaker 2011, 2013, 2015) is that paying closer attention to *all* of our responsibility-responses reveals that they have subsets that actually track three distinct agential capacities, capacities that are quite independent of one another and so do reflect three distinct *types* of responsibility. Paradigmatic forms of admiration and disdain fittingly respond to the quality of character of their agential targets, and so they require the capacities for character that include cares (emotional dispositions) and commitments (evaluative stances). This is *attributability*. Paradigmatic forms of pride and regret fittingly respond to the quality of judgment of their targets, and so presume various kinds of rational and evaluative capacities. As these capacities enable one in principle to answer a key responsibility-associated question—"Why did you do *that* (instead of *this*)?"—they generate *answerability*. Finally, paradigmatic forms of gratitude and anger fittingly respond to their target's quality of

[30] These are also what we might think of as *ethical* responses, not the responses associated with a much narrower kind of duty- or claim-based morality.

regard for others, itself a function of the capacity for empathy. This type of responsibility is labeled *accountability*. On my view, these are fully independent types of responsibility; none of them requires any other. Indeed, this tripartite theory best explains, I argue, the ambivalence felt toward so-called marginal agents (those with various psychological, personality, or mood disorders), who are in fact responsible in some ways but not others, and so warrant some subsets of our responsibility-responses but not others (in stark contrast to most previous monistic theories of responsibility which excluded marginal agents from the responsibility domain altogether).

Several theorists have come round to admitting that there are some important distinctions being made here, but they tend to have one of three resisting reactions in response. First, some deny that the distinctions do enough to get us to doubt the monistic nature of responsibility, as they can be recognized and still incorporated under the rubric of the right theory (see, e.g., Smith 2012, 2015; Talbert 2012b). For example, one might claim that the most fundamental unifying feature in all cases of responsibility is *answerability*, the ability in principle to answer the demand, "Why did you do that?" The capacity for answerability, it is thought, is simply the rational capacity to form evaluative judgments about the worth of reasons, and it is what undergirds and grounds *all* of our basic responsibility responses. So to the extent that there are distinctions in our set of responsibility-responses, they are distinctions only with respect to the different normative *domains* in which the answerability demand might be made and met. Responses to moral failures may well be different than those to aesthetic or prudential failures, but they are all matters of answerability, and they all target the same general rational capacities, so there is no reason to believe there is more than one type of responsibility.[31]

A second strategy is to allow for the distinctions to be compatible with a monistic theory of responsibility by denying that some of the responses are tracking *responsibility*. The near-universal way of doing so is to adopt accountability as the one "real" type of responsibility, and then to deny that attributability (on both Watson's and my schema) is actually a type of responsibility. Instead, for these theorists, it is thought to be merely a type of moral or ethical agency. One can be a moral agent, capable of doing morally good and bad things, without being a morally *responsible* agent, though.[32]

[31] The best version of this response has been developed by Angela Smith over the course of many articles, including Smith (2005, 2008, 2012, 2015).

[32] For the distinction between moral agency and morally responsible agency, see McKenna (2012, 9–14). For versions of the position that attributability, e.g., is not a form of responsible agency, see Vargas (2013, 103–4), and Driver (2015).

The third, and most popular, response these days is to say the following: "Yes, I recognize that attributability is a distinct type of responsibility (and perhaps answerability too), but the type of responsibility that *I'm* interested in, and will mostly or exclusively discuss, is accountability."[33] This is to admit pluralism but deny its relevance. It is a strategy typically deployed by those with metaphysical hankerings, as they are still most interested in the form of responsibility that ostensibly implicates sanctions and harsh treatments and so cries out for moral justification. Attributability—which is just about how people *are* and aptly generates only aretaic predication and emotions in response—does not implicate such treatments, and so is taken to be irrelevant to the issues that *matter*, to the issues that people have been debating in the free will literature for nearly 2000 years.

I find this last response rather baffling. The whole point of Watson's original pluralistic (or dual aspect) theory was to reveal and defend just *how* important and deep attributability is as a face of responsibility. It goes to the heart of practical agency, after all, to the source of who we are as we make our imprints in the world and on each other, and so it is hard to think of anything deeper or more important in our interpersonal interactions. We reorganize our entire lives sometimes in response to people's character traits and manifestations thereof, and we can locate one another in the world via those manifestations. Accountability, on Watson's view, adds only a kind of control to this mix, but our capacity for control adds *nothing* to the agential character being manifested by accountable agents, and it is other people's agential character which is the source of most of our own agential adjustments or engagements in response.[34]

This is a very serious fault line. Some may still think it "merely" a conceptual fault line, a boring semantic dispute over what counts as "responsibility." This is false. While it is true that there is a conceptual dispute involved here, it is not "merely" such a dispute. It is more fundamentally a tug of war over how to construe the relation between how we (ought to) respond attitudinally to one another in virtue of agential expressions and how we (ought to) *treat* one another in virtue of agential expressions (with the addition of control, *maybe*). This dispute goes not only to how

[33] There are more or less explicit versions of this view. McKenna says it straightforwardly in 2012, 8. Clarke (2014, 111–15) implies it. And Pereboom (2014), while admitting pluralism about responsibility, is famous for focusing heavily on, by strenuously arguing *against* the metaphysical plausibility of, accountability. Many others, in private communication, have told me that they hold this third view, even if they have not explicitly said it in print.

[34] This is a point recognized by McKenna and Van Schoelandt (2015), which makes it all the more surprising that McKenna still leans in the direction of adopting this third view against the full-on pluralist proposal (private correspondence).

unimpaired agents should respond to and treat each other but also to how those with various disorders and disabilities are to be *responded to* and *treated*, and to whether they are to be incorporated into the responsibility community at all. This is an extremely serious issue, as those who are excluded from the responsibility community are also often excluded from a wide variety of goods distributed only within that community, including the goods of recognition, respect, and interpersonal emotional engagement. If there would be great unfairness in treating the nonaccountable as though they were accountable, there would also be (and is) significant moral cost to treating those marginal agents who may be responsible (albeit in nonaccountable ways) as though they were not.[35]

FINAL REFLECTIONS

Some may contend that a few of the fault lines discussed above are not foundational disputes. Perhaps, for example, a freedom requirement of some sort follows trivially from other considerations, like the sorts of expectations or demands that are implicated when we hold one another responsible.[36] And some will surely contend that there are foundational fault lines that I have missed. Perhaps, for instance, there are fault lines over a basic form of personal responsibility, that is, over what the conditions are for actions or attitudes to be attributable to one in order to be eligible for the range of responsibility-responses we have discussed.[37] Or perhaps there are fault lines over what counts as the right object of direct responsibility, that is, what thing(s) we can be directly responsible for, where the contenders are actions, attitudes, and perhaps even perceptual states (e.g., empathy).[38]

My list of fault lines may be somewhat idiosyncratic, as they represent many of my own foundational interests. But the disputes I *have* noted are in fact traceable to Strawson, whose work was revolutionary, and who remains the source of much contemporary contention. At this point, it is unclear whether the fault lines he introduced are bridgeable, or whether one must simply pick a side and stick with it, perhaps waiting for the other side to die

[35] For much more on this last point, see my unpublished paper "Disordered, Disregarded, Disabled, Dismissed: Exemptions and Immorality."

[36] For example, Scanlon (1998, ch. 6) maintains that freedom simply falls out of the requirements of giving people the fair opportunity to avoid various costs involved in bearing the burdens of their choices when things go wrong (see also Hieronymi, forthcoming). Thanks to Michael McKenna for drawing my attention to this point.

[37] This has been Angela Smith's focus in recent years. See, e.g., Smith (2012, 2015).

[38] See, e.g., the exchange between Smith (2005, 2008), Levy (2005), and; see also Portmore (2019), as well as my unpublished paper "Empathic Control?"

off (as in a paradigm shift in science). But these contentious issues are part of a relatively new field struggling to establish its identity. It currently has many States of the *Questions*, and we must recognize and grapple directly with its fault lines in order to enable responsibility theory to grow up, perhaps to become its own substantial and distinctive area of philosophical inquiry along the lines of other well-established areas of philosophy like ethics, metaphysics, and epistemology.[39]

REFERENCES

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

———. 2006. *Merit, Meaning, and Human Bondage: An Essay on Free Will*. Princeton, NJ: Princeton University Press.

Arpaly, Nomy, and Timothy Schroeder. 2014. *In Praise of Desire*. Oxford: Oxford University Press.

Bennett, Jonathan. 1980. "Accountability." In *Philosophical Subjects: Essays Presented to P. F. Strawson*, edited by Zak van Straaten, 14–47. Oxford: Clarendon Press.

Bjornsson, Gunnar. 2017. "Explaining (Away) the Epistemic Condition on Moral Responsibility." In *Responsibility: The Epistemic Condition*, edited by Philip Robichaud and Jan Willem Wieland, 146–62. Oxford: Oxford University Press.

Brink, David O., and Dana K. Nelkin. 2013. "Fairness and the Architecture of Responsibility." *Oxford Studies in Agency and Responsibility* 1: 284–313.

Campbell, C. A. 1957. "Has the Self 'Free Will'?" In *On Selfhood and Godhood*, edited by C. A. Campbell, 158–79. London: George Allen and Unwin.

Carlsson, Andreas Brekke. 2017. "Blameworthiness as Deserved Guilt." *The Journal of Ethics* 21: 89–115.

———. 2019. "Shame and Attributability." *Oxford Studies in Agency and Responsibility* 6: 112–39.

Chisholm, Roderick. 1964. *Human Freedom and the Self: The Lindley Lecture*. University of Kansas.

Clarke, Randolph. 2005. "An Argument for the Impossibility of Moral Responsibility." *Midwest Studies in Philosophy* 29: 13–24.

[39] I am very fortunate to work in a field with generous, kind, and insightful fellow travelers. I wrote this essay after lengthy discussions over a few years with several of my colleagues and friends in the field: Randy Clarke, Andrew Eshleman, Michael McKenna, Dana Nelkin, Sandy Reiter, Angie Smith, Chandra Sripada, Matt Talbert, and Manuel Vargas. While many of the issues discussed in this paper are articulations of fault lines between members of this group as responsibility theorists, there are, I'm happy to say, no fault lines between us personally. I'm deeply grateful to them for their insights, and I'm especially grateful here to Randy and Angie for their written comments on an earlier draft of this essay. I also thank an anonymous referee for helpful remarks on the penultimate draft. Finally, I'm grateful to Remy Debes for urging me to write this piece, as well as for his patience in waiting for it to be written.

———. 2013. "Some Theses on Desert." *Philosophical Explorations* 16: 153–64.
———. 2014. *Omissions*. New York: Oxford University Press.
———. 2016. "Moral Responsibility, Guilt, and Retributivism." *The Journal of Ethics* 20: 121–37.
Clarke, Randolph, and Piers Rawling. Manuscript. *Reasons to Feel Guilty*. Florida State University.
Coates, D. Justin, and Neal A. Tognazzini, eds. 2013. *Blame: Its Nature and Norms*. New York: Oxford University Press.
D'Arms, Justin, and Daniel Jacobson. 2000. "The Moralistic Fallacy: On the 'Appropriateness' of the Emotions." *Philosophy & Phenomenological Research* 61: 65–90.
D'Arms, Justin, and Daniel Jacobson. 2014. "Wrong Kinds of Reason and the Opacity of Normative Force." *Oxford Studies in Metaethics* 9: 215–44.
Darwall, Stephen. 2006. *The Second-Person Standpoint*. Cambridge, MA: Harvard University Press.
Driver, Julia. 2015. "Appraisability, Attributability, and Moral Agency." In *The Nature of Moral Responsibility*, edited by Randolph Clarke, Michael McKenna, and Angela M. Smith, 157–74. New York: Oxford University Press.
Duff, Antony. 2007. *Answering for Crime*. Oxford: Hart Publishing.
Duggan, A. P. 2018. "Moral Responsibility as Guiltworthiness." *Ethical Theory and Moral Practice* 21: 291–309.
Feinberg, Joel. 1970. *Doing and Deserving*. Princeton, NJ: Princeton University Press.
Fischer, John. 2007. "Compatibilism." In *Four Views on Free Will*, edited by J. M. Fischer, R. Kane, D. Pereboom, and M. Vargas, 44–84. Oxford: Blackwell Publishers.
Fischer, John, and Mark Ravizza. 1993. "Introduction." In *Perspectives on Moral Responsibility*, edited by John Fischer and Mark Ravizza, 1–41. Ithaca, NY: Cornell University Press.
Fischer, John, and Mark Ravizza. 1998. *Responsibility and Control*. Cambridge: Cambridge University Press.
Frankfurt, Harry (1988). *The Importance of What We Care About*. Cambridge: Cambridge University Press.
Fricker, Miranda. 2016. "What's the Point of Blame? A Paradigm Based Explanation." *Noûs* 50: 165–83.
Ginet, Carl. 1966. "Might We have No Choice?" In *Freedom and Determinism*, edited by Keith Lehrer, 87–104. New York: Random House.
Glover, Jonathan. 1970. *Responsibility*. London: Routledge and Kegan Paul.
Gorman, August. 2019. "The Minimal Approval Account of Attributability." *Oxford Studies in Agency and Responsibility* 6: 140–64.
Haji, Ishtiyaque. 1998. *Moral Appraisability: Puzzles, Proposals, and Perplexities*. New York: Oxford University Press.
Harman, Elizabeth. 2011. "Does Moral Ignorance Exculpate?" *Ratio* 24: 443–68.
Hieronymi, Pamela. 2004. "The Force and Fairness of Blame." *Philosophical Perspectives* 18: 115–48.
———. 2014. "Reflection and Responsibility." *Philosophy & Public Affairs* 42: 3–41.
———. 2019. "I'll Bet You Think This Blame is About You." *Oxford Studies in Agency and Responsibility* 5: 60–87.

———. Forthcoming. "Fairness, Sanctions, and Condemnation." *Oxford Studies in Agency and Responsibility* 7.

Kane, Robert (1996). *The Significance of Free Will*. Oxford: Oxford University Press on Demand.

Lenman, James. 2006. "Compatibilism and Contractualism: The Possibility of Moral Responsibility." *Ethics* 117: 7–31.

Levy, Neil. 2005. "The Good, the Bad, and the Blameworthy." *Journal of Ethics & Social Philosophy*, 1.

———. 2011. *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford: Oxford University Press.

List, Christian. 2019. *Why Free Will is Real*. Cambridge, MA: Harvard University Press.

Macnamara, Coleen. 2013. "Taking Demands Out of Blame." In Coates and Tognazzini, 2013a, 141–61.

Mason, Elinor. 2019. *Ways to be Blameworthy*. Oxford: Oxford University Press.

McGeer, Victoria. 2013. "Civilizing Blame." In Coates and Tognazzini, 2013a, 162–88.

———. 2014. "P. F. Strawson's Consequentialism." *Oxford Studies in Agency and Responsibility*: 64–92.

McKenna, Michael. 2006. "Collective Responsibility and the Agent Meaning Theory." *Midwest Studies in Philosophy* 30: 16–34.

———. 2012. *Conversation and Responsibility*. New York: Oxford University Press.

———. 2013. "Directed Blame and Conversation." In Coates and Tognazzini, 2013a, 119–40.

McKenna, Michael, and Chad Van Schoelandt. 2015. "Crossing a Mesh Theory with a Reasons-Responsive Theory: Unholy Spawn of an Impending Apocalypse or Love Child of a New Dawn?" In *Agency, Freedom, and Moral Responsibility*, edited by Andrei Buckareff, Carlos Moya, and Sergi Rosell, 44–64. London: Palgrave Macmillan.

Nelkin, Dana. 2008. "Responsibility and Rational Abilities: Defending an Asymmetrical View." *Pacific Philosophical Quarterly* 89: 497–515.

———. 2011. *Making Sense of Freedom and Responsibility*. Oxford: Oxford University Press.

Nichols, Shaun. 2015. *Bound: Essays on Free Will and Responsibility*. Oxford: Oxford University Press.

Pereboom, Derk. 2001. *Living Without Free Will*. Cambridge: Cambridge University Press.

———. 2014. *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.

Portmore, Douglas W. 2019. "Control, Attitudes, and Accountability." *Oxford Studies in Agency and Responsibility* 6: 7–32.

Railton, Peter. 1984. "Alienation, Consequentialism, and the Demands of Morality." *Philosophy & Public Affairs* 13: 134–71.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: The Belknap Press of Harvard University Press.

Rosen, Gideon. 2004. "Skepticism about Moral Responsibility." *Philosophical Perspectives* 18: 295–313.

———. 2015. "The Alethic Conception of Responsibility." In *The Nature of Moral Responsibility*, edited by Randolph Clarke, Michael McKenna, and Angela M. Smith, 65–88. Oxford: Oxford University Press.

Russell, Paul. 2017. *The Limits of Free Will*. New York: Oxford University Press.

Sartorio, Carolina. 2016. *Causation and Free Will*. Oxford: Oxford University Press.
Scanlon, T. M. 1988. "The Significance of Choice." In *The Tanner Lectures on Human Values*, vol. 8, edited by Sterling McMurrin, 149–216. Salt Lake City: University of Utah Press.
———. 1998. *What We Owe to Each Other*. Cambridge, MA: The Belknap Press of Harvard University Press.
———. 2008. *Moral Dimensions*. Cambridge, MA: Harvard University Press.
———. 2013. "Giving Desert its Due." *Philosophical Explorations* 16: 101–16.
Sher, George. 2006. *In Praise of Blame*. Oxford: Oxford University Press.
Shoemaker, David. 2011. "Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility." *Ethics* 121: 602–32.
———. 2013. "Qualities of Will." *Social Philosophy & Policy* 30: 95–120.
———. 2015. *Responsibility from the Margins*. Oxford: Oxford University Press.
———. 2017. "Response-Dependent Responsibility; Or, a Funny Thing Happened on the Way to Blame." *Philosophical Review* 126: 481–527.
———. 2019. "Hurt Feelings." *The Journal of Philosophy* CXVI: 125–48.
———. Forthcoming. "Response-Dependent Theories of Responsibility." In *The Oxford Handbook of Moral Responsibility*, edited by Dana Nelkin and Derk Pereboom. Oxford: Oxford University Press.
Shoemaker, David, and Manuel Vargas. 2019. Moral Torch Fishing: A Signaling Theory of Blame. *Nous*. https://doi.org/10.1111/nous.12316.
Smart, J. J. C. 1973. "An Outline of a System of Utilitarian Ethics." In *Utilitarianism: For and Against*, edited by J. C. Smart and Bernard Williams, 3–76. London: Cambridge University Press.
Smith, Angela. 2005. "Responsibility for Attitudes: Activity and Passivity in Mental Life." *Ethics* 115: 236–71.
———. 2008. "Control, Responsibility, and Moral Assessment." *Philosophical Studies* 138: 367–92.
———. 2012. "Attributability, Answerability, and Accountability. In Defense of a Unified Account." *Ethics* 122: 575–89.
Smith, Angela M. 2013. "Moral Blame and Moral Protest." In Coates and Tognazzini, 2013a, 27–48.
Smith, Angela. 2015. "Responsibility as Answerability." *Inquiry* 58: 99–126.
———. 2019. "Who's Afraid of a Little Resentment?" *Oxford Studies in Agency and Responsibility* 6: 85–111.
Sommers, Tamler. 2012. *Relative Justice*. Princeton, NJ: Princeton University Press.
Sripada, Chandra. 2016. "Self-Expression: A Deep Self Theory of Moral Responsibility." *Philosophical Studies* 173: 1203–32.
Stace, Walter T. 1953. *Religion and the Modern Mind*. New York: MacMillan.
Strawson, Galen. 1994. "The Impossibility of Moral Responsibility." *Philosophical Studies* 75: 5–24.
Strawson, P. F. 1962. "Freedom and Resentment." *Proceedings of the British Academy* 48: 1–25. Reprinted in, and all page references from, Gary Watson, ed., 2003. *Free Will*, 2nd ed. Oxford: Oxford University Press, pp. 72–93.
Tadros, Victor. 2005. *Criminal Responsibility*. Oxford: Oxford University Press.

Talbert, Matthew. 2012a. "Moral Competence, Moral Blame, and Protest." *The Journal of Ethics* 16: 89–109.

———. 2012b. "Accountability, Aliens, and Psychopaths: A Reply to Shoemaker." *Ethics* 122: 562–74.

———. 2013. "Unwitting Wrongdoers and the Role of Moral Disagreement in Blame." *Oxford Studies in Agency and Responsibility* 1: 225–45..

Todd, Patrick. 2016. "Strawson, Moral Responsibility, and the 'Order of Explanation': An Intervention." *Ethics* 127: 208–40.

Tsai, George. 2017. "Respect and the Efficacy of Blame." *Oxford Studies in Agency and Responsibility* 4: 248–75.

Van Inwagen, Peter. 1983. *An Essay on Free Will*. Oxford: Clarendon Press.

Vargas, Manuel. 2013. *Building Better Beings*. Oxford: Oxford University Press.

Wallace, R. J. 1994. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

———. 2011. "Dispassionate Opprobrium: On Blame and the Reactive Sentiments." In Wallace, Kumar, and Freeman, 2011, 348–72.

Watson, Gary. 2004. *Agency and Answerability*. Oxford: Oxford University Press.

———. 2014. "Peter Strawson on Responsibility and Sociality." *Oxford Studies in Agency and Responsibility* 2: 15–32.

———. 2019. "Second Thoughts." *Oxford Studies in Agency and Responsibility* 5: 214–39.

Wiggins, David. 1973. "Towards a Reasonable Libertarianism." In *Essays on Freedom and Action*, edited by Ted Honderich, 31–62. London: Routledge and Kegan Paul.

Williams, Bernard. 1995. *Making Sense of Humanity*. Cambridge: Cambridge University Press.

Wolf, Susan. 1990. *Freedom within Reason*. Oxford: Oxford University Press.

———. 2011. "Blame, Italian Style." In Wallace, Kumar, and Freeman, 2011, 332–47.

Zimmerman, Michael. 1988. *An Essay on Moral Responsibility*. Totowa, NJ: Rowman and Littlefield.