[Forthcoming in *Law and Philosophy*]

Holding Responsible and Taking Responsibility

STEPHEN BERO
University of Surrey School of Law

Abstract: In matters of responsibility, there are often two sides to the transaction: one party who *holds* another responsible, and the other who (ideally) *takes* responsibility for her conduct. The first side has been closely scrutinized in discussions of the nature of responsibility, due to the influential Strawsonian conjecture that an agent is responsible if and only if it is (in some sense) appropriate to hold her responsible.

This preoccupation with holding responsible—with its focus on the second-personal perspective and on responses like blame—contrasts with a relative neglect of the perspective of the agent and the role that she has to play by taking responsibility. I aim to show that this neglect is undeserved—that taking responsibility is both distinct in character from holding responsible and fundamentally important in its own right. I develop a conception of taking responsibility that reveals an under-explored dimension of our responsibility practices.

In matters of responsibility, there are often two sides to the transaction: one party who *holds* another responsible, and the other who (ideally) *takes* responsibility for her conduct. The first side of this transaction has been closely scrutinized in discussions of the nature of responsibility, due in large part to the Strawsonian conjecture that an agent is responsible if and only if it is (in some sense) fitting or appropriate to hold her responsible—and further that *what it is* for an agent to be responsible perhaps *just is* for it to be appropriate to hold her responsible.[1] According to this way of thinking, to understand what it is to hold someone responsible, and when it is appropriate to do so, is to understand what it is, in at least one central sense, to be responsible.

This preoccupation with holding responsible—with its focus on the second- personal perspective, and on responses like blame, reproach, and sanction—contrasts with a relative neglect, in discussions of the nature of responsibility, of the perspective of the agent and the role that she has to play by taking responsibility. The latter perspective has, however, received considerably more attention in discussions of downstream topics like punishment, moral repair, and the moral emotions, which are seen to encompass a wider range of responses and activities than the narrower (and perhaps more fundamental) question of what it is to be responsible.

[1] The conjecture originates in Peter Strawson's "Freedom and Resentment," in Gary Watson (ed.), *Free Will* (Oxford and New York: Oxford University Press, 2003, 2nd ed.), pp. 72-93.

For instance, punishment theorists examine the interactions between an agent's responses to her own misconduct (remorse, reparation, etc.) and the appropriateness of sanctions[2]; reconciliation theorists offer accounts of how the agent can actively participate in processes of apology, atonement, and relational repair[3]; and theorists of the emotions explore the complex range of attitudes (guilt, shame, remorse, (agent-)regret, pride, etc.) that we can have towards our own conduct.[4] Given the range and richness of this work, it is striking—or at least, by the end of this paper I hope it will be striking—that efforts to understand the nature of responsibility in terms of the appropriateness of certain attitudes and responses have focused on the second-personal attitudes and responses associated with holding responsible (blame, reproach, etc.), largely to the exclusion of the attitudes and responses associated with taking responsibility.

I aim to show that this neglect of the agent's perspective is undeserved—that taking responsibility is both distinct in character from holding responsible and fundamentally important in its own right. To this end I will draw upon an idea that Joseph Raz has applied to the puzzling case of responsibility for inadvertent conduct, but I will repurpose this idea and apply it to ordinary deliberate conduct rather than inadvertence. In doing so, I will develop a conception of taking responsibility that reveals an under-explored dimension of our responsibility practices.

If this effort succeeds, there will be two significant upshots for theoretical inquiry into responsibility and its associated responses and practices. The first and more straightforward is that taking responsibility turns out to be independently important and to deserve more focused theoretical attention than it has generally received. The second is that we have reason to suspect that efforts to understand the nature and conditions of responsibility *solely* in terms of the attitudes and responses associated with holding responsible—as productive and illuminating as these efforts have been—cannot help but miss something vital. This will particularly be so if it should turn out, as I will briefly suggest in closing, that holding responsible and taking responsibility are not merely distinct and complementary but interactive and interdependent patterns

---

[2] *See*, *e.g.*, R.A. Duff, *Punishment, Communication, and Community* (Oxford and New York: Oxford University Press, 2001); John Tasioulas, "Punishment and Repentance," *Philosophy* 81(2) (2006): pp. 279-322; Christopher Bennett, *The Apology Ritual: A Philosophical Theory of Punishment* (Cambridge and New York: Cambridge University Press, 2008); Hannah Maslen, *Remorse, Penal Theory and Sentencing* (Oxford and Portland, Oregon: Hart Publishing, 2015).

[3] *See*, *e.g.*, Margaret Urban Walker, *Moral Repair: Reconstructing Moral Relations After Wrongdoing* (Cambridge and New York: Cambridge University Press, 2006); Nick Smith, *I Was Wrong: The Meanings of Apologies* (Cambridge and New York: Cambridge University Press, 2008); Linda Radzik, *Making Amends: Atonement in Morality, Law, and Politics* (Oxford and New York: Oxford University Press, 2009).

[4] *See*, *e.g.*, Bernard Williams, "Moral Luck," in *Moral Luck: Philosophical Papers 1973-1980* (Cambridge and New York: Cambridge University Press, 1982), pp. 20-39; John Deigh, "Love, Guilt, and the Sense of Justice," in *The Sources of Moral Agency* (Cambridge and New York: Cambridge University Press, 1996), pp. 39-64; Gabriele Taylor, *Pride, Shame, and Guilt: Emotions of Self-Assessment* (Oxford and New York: Clarendon Press, Oxford University Press, 1985).

of response. For in that case, responsibility ought to be understood as grounded in a more complex matrix of concerns and practices than has often been supposed.

Two clarifications are in order at the outset. First, while I am broadly sympathetic to the Strawsonian project of seeking to understand the nature of responsibility in terms of the attitudes and responses that make up our responsibility practices, my aim here is neither to offer a defense of that project nor directly to consider the nature of responsibility. I will not here defend any view, for instance, on the question of whether the attitudes and responses associated with responsibility are better understood as somehow constitutive of responsibility, or rather merely as licensed by an independent status of responsibility.[5] My aim is rather to shed light on a neglected aspect of our responsibility practices. This will naturally be of interest to Strawsonians who consider our responsibility practices to be the very essence of responsibility; but I hope it may also be of broader interest, given the importance of our responsibility practices in our moral and social lives.

Second, while I will deny both that holding responsible is more fundamental than taking responsibility, and that taking responsibility is to be understood in terms of holding responsible, I do not mean to endorse the inverse claims that taking responsibility is more fundamental than holding responsible, or that holding responsible is to be understood in terms of taking responsibility. My contention is rather that these two patterns of response are distinct and irreducible to one another, and that neither should be accorded theoretical priority over the other.

1 Taking responsibility and holding oneself responsible

First we need a clearer sense of the difference between 'holding responsible' and 'taking responsibility.' While these expressions are suggestive in themselves, I am proposing to use them as terms of art to capture a distinction that requires further explanation, especially as it can easily be overlooked. So far I have indicated that holding responsible is characteristically a second-personal way of responding to another agent's conduct (as in blaming or reproaching[6]), whereas what I am calling taking

[5] For discussion of this question *see*, *e.g.*, Gary Watson, "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme," in *Agency and Answerability* (Oxford and New York: Oxford University Press, 2004), pp. 219-59, 222; R. Jay Wallace, *Responsibility and the Moral Sentiments* (Cambridge, Massachusetts and London: Harvard University Press, 1994), pp. 1, 85-95.

[6] Or perhaps praising, crediting, and feeling gratitude, among a variety of other positive responses. There is an active controversy in the responsibility literature concerning whether praise, credit, and other positive responses count as ways of 'holding responsible,' or whether only (some limited set of) negative responses (often identified as resentment, indignation, and guilt) have that status. *See*, *e.g.*, Wallace, *supra* note 5, p. 61; Coleen Macnamara, "Holding Others Responsible," *Philosophical Studies* 152 (2011): pp. 81-102, 89; Gary Watson, "Peter Strawson on Responsibility and Sociality," in David Shoemaker & Neal A. Tognazzini (eds.), *Oxford Studies in Agency and Responsibility, Volume 2: "Freedom and Resentment" at 50* (Oxford and New York: Oxford University Press, 2014), pp. 15-32, 28. While I incline towards the view that for many purposes it makes sense to count positive responses like praise and credit among the ways in which we hold one another responsible, there are also important differences between the negative and positive responses that make the question a complex one. As I do not

responsibility is a first-personal way in which the agent can respond to her own conduct (as in feeling sorry or apologizing[7]). But of course this is just a first step.

To take the next step in drawing out this distinction, it is helpful to consider a doubt that immediately suggests itself: Perhaps there is fundamentally only one kind of response here, because taking responsibility is not in its essence distinct from holding responsible, but rather just boils down to *holding oneself responsible*. That is, perhaps the relevant first-personal responses are not different in kind from the second-personal ones (blaming, reproaching, etc.), but are rather simply self-directed versions of them.

This thought is encouraged by the Strawsonian idea that holding someone responsible is essentially a matter of having the sort of emotional reaction towards them that comes in the personal, vicarious, and reflexive versions that Peter Strawson labeled resentment, indignation, and guilt, respectively—what he famously called the "reactive attitudes."[8] According to a simple version of this idea, my holding you responsible for your treatment of me is fundamentally a matter of (or involves, or is somehow closely connected to) my resenting you; my holding you responsible for your treatment of a third party is a matter of my feeling indignant towards you—which is another way of saying that that I feel resentment, on behalf of the third party, towards you; and my holding myself responsible for my treatment of another is a matter of my feeling guilty—which is another way of saying that I feel resentment, on behalf of that other, towards myself. This way of carving up the phenomena suggests the hypothesis that to take responsibility is nothing other than to hold oneself responsible, in particular by undertaking the emotional burden of guilt and being disposed to respond in the ways that are associated with a sense of guilt.

But this, I contend, is mistaken. The goal of this section is thus to demonstrate that there is a fundamental distinction to be made between holding oneself responsible (in the Strawsonian sense just described) and the pattern of response that I am calling taking responsibility. The demonstration will proceed by illustrating a range of familiar attitudes and responses that we can, and at least sometimes do, take towards our own conduct, and by introducing some principled distinctions among those responses. Proceeding in this way will reveal a distinct set of attitudes and responses whose character and significance as a class have tended to be overlooked; along the way we will also designate a vocabulary to describe this relatively neglected category.

A crucial reason to think that we have here two distinct patterns of response is that among the familiar ways in which we sometimes respond to our own conduct we

---

intend to defend a view on this question here, I will focus throughout on negative responses.

[7] Or perhaps feeling pride and claiming credit; but for the reasons discussed in the previous footnote, the focus here will be on negative responses.

[8] In fact, Strawson conceived the reactive attitudes as a broader category, including in addition responses like hurt feelings and shame, as well as positive responses like gratitude, forgiveness, and love, *see supra* note 1, pp. 75, 79, 83-85. In the responsibility literature it has become common, due to the influence of Wallace and like-minded theorists working within the Strawsonian tradition, to limit discussion to negative attitudes, and specifically to resentment, indignation, and guilt. *See* Wallace, *supra* note 5, pp. 62-73. On the question of whether positive responses should be counted among the ways in which we hold one another responsible, *see supra* note 6.

can differentiate two sorts of emotional reactions that are distinguished by their characteristic foci or objects of concern. On the one hand are the emotional reactions that you, the active party, can have towards yourself *as agent*[9]; these include guilt, in the standard Strawsonian sense, and more generally self-directed anger, blame, frustration, and the like (perhaps also disappointment, contempt, and a range of other responses). These are self-directed versions of the attitudes that we take when holding another responsible, and they find natural expression in self-directed versions of the dispositions and actions that are associated with holding another responsible, including reproach, sanction, and so on. The agent's primary focus in these responses is on herself, her choices, and what she did.

On the other hand are the emotional reactions that you, the active party, can have towards another person *as someone affected by your conduct* and *as a party to the relationship* that you stand in to them. Here anger, blame, and the like are clearly out of place. What is called for instead are what we might call (somewhat stipulatively, but I think without unduly straining familiar usage) *contrition*—or more colloquially, *feeling sorry*—towards the person your conduct has affected. Associated with this reaction are dispositions to apologize, to 'undo' the wrong, to make amends, and so on. The agent's primary focus in these responses is on the other party, how her conduct has affected them, and the significance of this for their relations.

The literature on blame, guilt, and other agent-focused responses is vast; by contrast, discussion of this second set of responses is sparse enough that I have resorted to terms—contrition, feeling sorry—that have no established theoretical usage. (I will come in a moment to the related phenomena of agent-regret and remorse.) Yet feeling sorry is, I suspect, at least for many people no less familiar than feeling guilty. And that these two responses are entirely distinct is apparent from everyday situations in which it is possible to feel sorry for something you don't feel guilty for, and vice versa.

Suppose you make plans to see a friend who is very eager to spend time with you, after having cancelled several previous meetings (all with excellent reason). As luck would have it, when the appointed time comes another legitimate crisis arises and you have no choice but regretfully to cancel again. In this kind of case it is a possible and indeed I believe a familiar reaction not to feel guilty at all (you had excellent reason to cancel—there's nothing else you could reasonably have done), but at the same time to feel sorry towards your friend for letting her down again. And this feeling will tend to manifest itself in a disposition to apologize and offer explanations, a desire to 'make it up to her,' and so on, rather than in a disposition to reproach or sanction yourself. When you think about what happened, the focus will primarily be on your friend, your relationship, and the disappointment you have occasioned, rather than on yourself, your choices, or your conduct (though it is of course important that it was *your conduct* that occasioned this disappointment—in this way contrition is quite distinct from impersonal regret).

---

[9] The term 'agent' is used here not in any technical sense but simply to designate the active party—that is, the party whose conduct or activity is the occasion for the attitudes and responses in question. At this stage, for purposes of establishing the distinctness of the relevant types of response, we do not need to inquire further into the kind of agency or agential capacity that is necessary for the responses to be intelligible or appropriate.

The claim here is not, to be clear, that the response just imagined is necessarily the most fitting or appropriate one, nor that other possible responses (like feeling guilty) would be unfitting or inappropriate. Moreover, I mean neither to suggest that this response would be the most common or typical response in this scenario, nor to make any conceptual claim about what it means to feel 'sorry' or 'contrite.' My more modest aim is rather to isolate and identify a type of response that is possible and intelligible (and, I venture, familiar) in this sort of scenario, and one that is distinguished—by its characteristic focus or object of concern, and in other ways—from agent-focused responses like (Strawsonian) guilt or self-blame. And to have a vocabulary for speaking about the sort of response here distinguished, I propose stipulatively (but, I hope, without great strain or artificiality) to use the terms *contrition* or *feeling sorry*. What is at stake then are the possibility and the distinctiveness of the response described, not its fittingness or appropriateness (nor the fittingness or appropriateness of alternative possible responses)—the suggestion is that the imagined scenario is one in which we can recognize the possibility of feeling sorry (in the relevant sense) without feeling guilty.

Now consider a case in which you recklessly risk the safety of an indeterminate number of people, but through sheer good luck no harm occurs. On a night out you have several more cocktails than you planned, and decide to drive home when you should know better. On waking the next morning, you might feel guilty about what you've done ("What was I thinking? I could have killed someone!"), but without feeling sorry towards anyone—after all, no one was hurt, and the people whose safety you risked are unidentifiable. Your feeling will naturally manifest itself in a disposition, for instance, to reproach yourself, rather than to apologize or make amends (to whom?). When you think about what happened, the focus will primarily be on yourself, your choices, and what you did, rather than on any other person.

Again, the claim is not that this response would be more appropriate or more likely to occur than alternative possible responses. We could perhaps imagine, for instance, someone in this scenario focusing instead on the people (either in particular or in the abstract) whose safety she risked, and orienting her emotional response around concern for, or contrition towards, them. The claim is rather that the more self-focused pattern of response here described is possible and intelligible (and, I again venture, familiar).

For another example of feeling guilty without feeling sorry we can borrow a case imagined by R. Jay Wallace, in which a lapsed Catholic feels irrationally guilty for having recreational sex, despite no longer endorsing the rule that this guilt is based on.[10] While Wallace does not remark on this aspect of the case, we can note that this person need not also feel *sorry*. He might, for instance, feel (irrationally) angry at himself and disposed to reproach or to punish himself, but without being disposed to apologize to, or feel sorry towards, anyone at all.

The previous two examples are useful for isolating feelings of guilt from feelings of contrition because they involve no specific other person to whom it would obviously make sense to feel sorry. But it is quite possible to have only feelings of guilt even in cases where there is a readily identifiable victim. Suppose you allow yourself an intemperate remark addressed to a person whom you simply loathe. This person, you are convinced, deserved no better—indeed, much worse than what you said. And

[10] *See* Wallace, *supra* note 5, p. 43.

yet . . . you prefer to hold *yourself* to a higher standard. Far from feeling sorry or disposed to apologize, you might continue to take satisfaction in having given this person a bit of what they had coming, even as you blame yourself for your lack of self-control. In this way, you might feel guilty, be disposed to reproach yourself, and so on, despite not feeling at all sorry towards the target of your remark.

These cases thus distinguish two sorts of emotional reaction by their characteristic foci or objects of concern, among other features. As I have already noted, agent-focused responses like guilt have been extensively discussed in the literature on the nature of responsibility; other-focused responses like contrition have received less attention in that literature, notwithstanding their familiarity in everyday experience. But in neighboring discussions[11] we find treatments of two related phenomena that are useful for comparison: agent-regret and remorse.

The notion of agent-regret was coined by Bernard Williams to highlight the special concern that we have about our own actions, even when those actions are faultless or have a significance that is entirely beyond our control (as in his example of a driver who, through no fault of his own, runs over a child). Williams therefore conceives of agent-regret as an appropriate response to a very wide range of conduct (in contrast to remorse, which he conceives as a narrower subset of agent-regret that is restricted in its application to voluntary wrongdoing). At the same time, agent-regret's scope is narrower than that of ordinary impersonal regret; one of the earmarks that distinguishes it from impersonal regret, Williams suggests, is that a desire to compensate or make reparation can be a natural expression of agent-regret.[12]

Williams's reference to other-directed responses like recompense and reparation indicates an important kinship between agent-regret and my notion of contrition. There are also, however, at least two important differences. First, whatever overlap there may be between them, the notion of agent-regret was introduced and conceived to serve quite different theoretical purposes. Williams sought to bring to light the full breadth of the special concern that we have about our own actions; despite invoking reparation as a possible expression of agent-regret, he was not directly concerned with the distinction between different foci or objects of concern that is central to the way I am conceiving of contrition. The notion of contrition thus more clearly isolates responses that are other-focused in the distinctive way I have described. Second, and related to this, the notion of agent-regret appears to apply much more broadly than contrition. I am conceiving of contrition in opposition to agent-focused responses like guilt, but there is no indication that Williams saw any similar opposition between agent-regret and guilt; rather, the more natural interpretation of Williams is that he understood agent-regret to be a basic concern that we have for our own conduct that is present as an element in experiences of both remorse and guilt, as well as a range of other responses like disappointment in oneself, shame at what one has done, and indeed contrition. Agent-regret is thus best seen as a perhaps necessary element or condition of contrition, but not a sufficient one. We can experience agent-regret, in other words, without feeling sorry towards anyone—even if feeling sorry may always involve an element of agent-regret.

---

[11] Such as those referenced *supra* notes 2-4.

[12] Williams, *supra* note 4, p. 28-29. More specifically, Williams suggests that it is natural in agent-regret, as it is not in impersonal regret, to feel that one should offer compensation oneself, regardless of whether the damage is also covered by insurance.

Remorse, unlike agent-regret, has sometimes been conceived in opposition to guilt, and it is thus in discussions of remorse that we find the closest precedent to my notion of contrition or feeling sorry. In her influential treatment of the contrast between guilt and remorse, for instance, Gabriele Taylor suggests that "[t]he important feature of guilt is that the thought of the guilty concentrates on herself as the doer of the deed"; in remorse, by contrast, the agent's "concentration of thought is . . . not on the self . . . but is on her actions and their consequences. It is more outward-looking" than guilt.[13] It is characteristic of remorse, says Taylor, that "in feeling remorse the agent believes that she has done harm which she ought to try and repair," whereas in guilt there may be no thought of harm or reparation at all, but only of her own transgression.[14] For this reason, Taylor concludes that remorse, unlike guilt, is not an "emotion of self-assessment."[15]

In these respects, Taylor's conception of remorse is a close match with, and an important model for, my notion of contrition. But in other respects Taylor's remorse is a thicker notion and a more moralized one, and it accordingly has a narrower range of application. Remorse, Taylor says, unlike impersonal regret, "never implies acceptance" of what has been done—"[i]t is impossible to feel remorse and yet believe that overall it was right to act as one did"; she also takes remorse to be a moral emotion in the sense that it requires "an awareness, more or less developed, of moral distinctions, of what is right or wrong, honourable or disgraceful"; and she considers that the remorseful agent must see "himself as a responsible moral agent, and so sees whatever wrong he has done as an action (or omission) of his about the consequences of which he ought, if possible, to do something."[16] These further characterizations are representative of much of the theoretical literature on remorse, in that remorse is often treated as implying an all-things-considered desire to have acted otherwise, as involving complex moral judgments, as being itself a significant moral achievement, as being an important or even a sufficient condition for forgiveness or suspension of punishment, and so on.[17]

I have no objection to conceiving of remorse in these thick, moralized ways, but I intend contrition or feeling sorry to be a thinner notion that applies to a broader range of responses. For instance, I wish to include the way you might feel sorry towards your friend for cancelling your meeting, even when you see your conduct as morally permissible (even required), or indeed fail to conceive it in moral terms at all and see it simply as having hurt your friend's feelings. This thinner notion of contrition puts the basic contrast with guilt and related responses into clearer relief, and (as we will see) provides the foundation we will need to begin to develop a general conception of taking responsibility.

---

[13] Taylor, *supra* note 4, pp. 97, 100.

[14] *Ibid.* pp. 103-104.

[15] *Ibid.* p. 99. A broadly similar way of distinguishing between guilt and remorse can be found in Deigh, *supra* note 4, p. 48, though Deigh does not attribute to remorse the further features discussed below that are found in Taylor. Deigh does, however, ground remorse directly in feelings of love or identification in a way that distinguishes his conception of remorse from my conception of contrition.

[16] Taylor, *supra* note 4, pp. 99, 107.

[17] For versions of several of these claims in a recent and sophisticated discussion of remorse, *see* Maslen, *supra* note 2, especially pp. 5-12.

The notion of contrition or feeling sorry, and the contrast between this and agent-directed responses like guilt, is now perhaps coming into focus. But before moving on, one final observation in favor of distinguishing these two types of response concerns an important structural feature. As Strawson observed, the second-personal responses associated with holding responsible also have first- and third-personal versions: I can be angry or resentful towards myself for what I have done to you, just as you can be angry or resentful towards me (similarly, I can reproach or punish myself, just as you can reproach or punish me). But contrition, in contrast, is a first-personal response with no second- or third-personal analogue: You cannot feel sorry towards yourself *for the way I have treated you* in anything like the way in which I can feel sorry towards you (just as you cannot apologize or offer amends to yourself in anything like the way in which I can apologize or offer amends to you)—indeed, even trying to express this idea reveals its incoherence.

Now, with this distinction between guilt and contrition in hand, we are in a position to make the further distinction that I aim to draw between *holding (oneself) responsible* and *taking responsibility* in a very straightforward way. Whereas the characteristic emotional dimension of holding oneself responsible is feeling *guilty*, the characteristic emotional dimension of taking responsibility is feeling *sorry*. Guilt, and holding oneself responsible, are agent-directed; contrition or feeling sorry, and taking responsibility, are directed towards the person whom one's conduct has affected. Thus, the same cases that illustrate how someone can feel sorry but not guilty, and vice versa, also illustrate how it is possible both to hold oneself responsible without taking responsibility, and to take responsibility without holding oneself responsible (though of course it is common and often appropriate to do both). The person who misses an appointment with a friend for excellent reasons may take responsibility (with respect to her friend) without holding herself responsible; similarly, the lucky drunk driver may hold herself responsible without taking responsibility (with respect to anyone).

This confirms, as was previewed earlier, that I am giving somewhat special, stipulative senses to the terms *holding responsible* and *taking responsibility* (as to the terms *contrition* and *feeling sorry*), in order to draw attention to a distinction that can otherwise easily be missed. In ordinary parlance, these terms are used more loosely; we do not typically have occasion to distinguish between finer senses of responsibility and what I am calling the taking versus the holding of it.[18] Nonetheless, the examples we have considered highlight a significant distinction that I will use the terms *holding responsible* and *taking responsibility* to identify.

If the foregoing is correct, we now have reason to think that taking responsibility and holding (oneself) responsible are distinct, but we lack as yet a satisfactory

[18] Also contrast my sense of 'taking responsibility' with the sense at issue in David Enoch, "Being Responsible, Taking Responsibility, and Penumbral Agency," in Ulrike Heuer & Gerald Lang (eds.), *Luck, Value, and Commitment: Themes from the Ethics of Bernard Williams* (Oxford and New York: Oxford University Press, 2012), pp. 95-132, and "Tort Liability and Taking Responsibility," in John Oberdiek (ed.), *Philosophical Foundations of the Law of Torts* (Oxford and New York: Oxford University Press, 2014), pp. 250-71, in which 'taking responsibility' is conceived as a (possibly mental) action by which someone *assumes* responsibility for something that would otherwise not be their responsibility. This sense of 'taking responsibility,' consisting in the exercise of a kind of normative power, is not my topic.

conception of taking responsibility—of what it is, what it means, and why we do it. My goal in the rest of this paper is to sketch such a conception. But I will proceed indirectly, via a detour through Joseph Raz's novel and resourceful treatment of responsibility for inadvertence. By borrowing and repurposing a core element of Raz's account, I hope significantly to advance our understanding of taking responsibility.

2 Raz on responsibility for inadvertence

In the final two chapters of *From Normativity to Responsibility*,[19] Raz is concerned with what we might call the puzzle of responsibility for inadvertence. The puzzle arises in the following way. On the one hand, we often hold ourselves and others responsible for various inadvertent failures: forgetting an appointment, accidentally botching a procedure, and so on. On the other hand, this can seem dubious: It is (relatively) easy to see why we are 'on the hook' for successful conduct that turns out just the way we intended, but harder to see why we should be on the hook for conduct that results inadvertently from failures of our powers. In particular, responsibility for inadvertent conduct appears to violate a plausible condition that would limit responsibility to conduct over which we successfully exercise some sort of control, and thereby subjects us to a potentially troubling kind of moral luck.[20] There seems then to be at least some reason to think that we are on firmer ground in holding agents (including ourselves) responsible for conduct that is successfully guided by their powers than we are in holding them responsible for conduct that isn't.

Raz undertakes to show that this is wrong, and that we are often on firm ground in holding ourselves responsible for inadvertent conduct. While Raz is not attentive to the distinction drawn in the previous section between holding responsible and taking responsibility, in the course of developing his account he focuses on attitudes and responses that the agent takes towards herself *as agent*, rather than attitudes and responses that she takes towards *the person affected by her conduct*. For this reason Raz can only be read as offering an account of why we hold ourselves responsible, in my sense. Indeed, we could say that one of his significant innovations is to offer an explanation of our responsibility for inadvertent conduct in terms of why we hold ourselves responsible for such conduct.

According to the principle of responsibility that Raz defends, we are responsible for "conduct that is the result of the functioning, *successful or failed*, of our powers of rational agency, provided those powers were not suspended in a way affecting the action."[21] By "powers of rational agency" Raz means quite broadly all of the powers that we can call upon in the service of our activity as rational agents who form and execute intentions in normal ways. Our powers of rational agency thus include our sub-personal capacities of memory, attention, perception, bodily control, will power, and all of the other abilities that we regularly call upon to translate our intentions into conduct.[22]

---

[19] Joseph Raz, *From Normativity to Responsibility* (Oxford and New York: Oxford University Press, 2011).

[20] *See ibid.* p. 243. For discussion of the conflict between moral luck and a plausible control condition on responsibility, *see generally* David Enoch & Andrei Marmor, "The Case Against Moral Luck," *Law and Philosophy* 26 (2007): pp. 405-36.

[21] Raz, *supra* note 19, p. 231 (emphasis added).

[22] *See ibid.* pp. 2, 227.

Raz's principle, then, makes us responsible not only for conduct that results from the successful functioning of our powers of rational agency, but also for conduct that results from failures of those powers[23]; as a result, it makes us responsible for a wide range of conduct that results from lapses of memory, coordination, will power, and all the rest (including some conduct traditionally classified as 'negligent'). Thus, when you fail to pick up your child from school because you simply forgot; when you hit the car in front of you because your foot simply slipped off the brake; and so on, you are responsible for those inadvertent lapses, according to Raz.[24]

But these examples do not quite convey the distinctive breadth of this view of responsibility for inadvertent conduct.[25] Consider that one familiar position is that we can be responsible for an inadvertent lapse of memory, coordination, or the like, *provided* that the lapse is a manifestation of an objectionable character trait, attitude, or judgment. The idea is that someone can be responsible for forgetting a friend's birthday, for instance, if this is a manifestation of insufficient concern for her friend.[26] Another common thought is that an agent can be responsible for an inadvertent lapse of memory, coordination, or the like, *provided* that the lapse is causally traceable to a previous error on her part that was *not* inadvertent. For instance, an agent can be responsible for forgetting a birthday if she had previously decided not to bother marking the birthday in her calendar, consciously disregarding the risk that she might later forget. Raz's view goes further than either of these ideas: He says that someone who simply forgets a friend's birthday—her memory *just fails* her on this occasion, as memories sometimes do—is responsible for the lapse *even if* her concern for her friend and her prior conduct have been entirely satisfactory, or even exemplary.

In order to defend this broad view of our responsibility for inadvertence, Raz develops some new theoretical resources. He begins with the idea of a "domain of secure competence" whose boundaries are defined by the powers of rational agency that we "securely command." The idea is that within our domains of secure competence we can be confident that (barring some competence-defeating condition, like a seizure) if we undertake to act we will succeed; we are entitled to act without reflecting on the prospects for success.[27]

Any exercise of complex agency, Raz argues, presupposes the stable existence of a domain of secure competence. It is our confidence in our domains of secure competence that allows us, for instance, to deliberate, decide, and act without constantly second-guessing ourselves or going back to verify the soundness of our deliberations[28]; in the same way, we rely on our domains of secure competence to navigate our physical world without constantly questioning and verifying the nature and extent of our basic abilities to control our bodies and manipulate our surroundings. Without taking a

---

[23] With the important qualification, to be explained below, that we are only responsible for failures of our powers that fall within what Raz calls our "domain of secure competence." *See ibid.* pp. 244-45.

[24] *See ibid.* pp. 244, 267.

[25] On this distinctive breadth, *see also* Gary Watson, "Raz on Responsibility," *Criminal Law and Philosophy* 10 (2016): pp. 395-409, 399.

[26] This example is from Angela M. Smith, "Responsibility for Attitudes: Activity and Passivity in Mental Life," *Ethics* 115 (2) (2005): pp. 236-71, 236.

[27] *See* Raz, *supra* note 19, pp. 244-45, 268.

[28] *See ibid.* p. 250.

substantial domain of secure competence for granted, we would be unable to rely unreflectively on our various lower-level capacities in order to perform any complex action, and would be effectively paralyzed.

Raz further contends that because possession of a domain of secure competence is an essential component of complex human agency, it is also central to our self-conceptions and identities as agents. Our self-esteem, our self-respect, our pride (or shame) in who we are and what we can do, our sense of our own potential—and thus our projects and ambitions—"all these and various other self-directed attitudes and emotions depend in part on competence in using our faculties of rational agency."[29] To have a shorthand for the complex of self-directed attitudes that Raz has in mind, I will refer to them collectively as one's 'self-conception as an agent.' For Raz, our self-conceptions as agents comprise many (though not all[30]) of those self-directed attitudes that are most critical to our senses of who we are, of our place in the world, and of our own worth. And according to Raz's account, our self-conceptions as agents depend on establishing and maintaining the domains of secure competence that we claim for ourselves.

Because our domains of secure competence play a central role in our self-conceptions as agents, Raz argues that any failure to perform an act that the agent considers to be within her domain of secure competence has a special significance—it threatens to undermine her very self-conception as an agent:

> Failure to control conduct within our domain of secure competence threatens to undermine our self-esteem and our sense of who we are, what we are capable of, etc. We must react to it. We may conclude that we are no longer able securely to perform that kind of action. We have grown frail, our competence is diminishing. We come to recognize our limitations. Commonly this is not the case, and we do not allow it to be. We assert our competence by holding ourselves responsible for it. To disavow responsibility is to be false to who we are.[31]

When this kind of failure occurs, Raz says, "We must react to it," and there are two options: the agent can either revise her self-conception by conceding that her domain of secure competence is not as wide as she thought, or she can reassert her competence by *holding herself responsible* for the failure. (By this point it is clear that Raz's account is concerned with holding oneself responsible, in my sense, rather than taking responsibility. The agent's focus in his account is squarely on herself, her performance, and her self-conception as an agent, rather than on the affected party.)

This need to uphold our self-conceptions as agents, according to Raz's account, explains our practice of holding ourselves responsible for the results of failures of our powers of rational agency: "In acknowledging responsibility for actions due to our rational powers we are simply affirming that they are our secure rational powers."[32] In other words, by holding ourselves responsible we rebut the threat that failures of our powers of rational agency pose to our domains of secure competence, and thereby reaffirm our self-conceptions as agents. By the same token, "To disavow responsibility

---

[29] *Ibid.* pp. 245, 268.

[30] Other important elements of our self-understandings include, for instance, "gender or ethnicity and their social meanings," *ibid.* p. 239.

[31] *Ibid.* p. 245.

[32] *Ibid.*

is to be false to who we are," because it is equivalent to renouncing our own self-conceptions as agents.[33]

To be clear, for Raz holding myself responsible is something more than simply acknowledging that the conduct in question is *my* conduct. I have forgotten what I had for dinner ten years ago today. I am 'responsible' for this in the sense that it is a reflection of my admittedly imperfect memory. At the same time, however, there is no pressure for me to *hold myself responsible* for forgetting (blame myself, reproach myself, etc.), because remembering such a minor detail years later is beyond the domain of secure competence that I claim for myself. In this sense—the sense that interests Raz—I am *not* (to be held) responsible for forgetting.

Notably, Raz does not ever say that the need to rebut the threat to our self-conceptions as agents furnishes us with a *reason* to hold ourselves responsible, or otherwise state that the connection between holding ourselves responsible and the need to rebut the threat is a rational one. His purposes may not require it to be, if I understand correctly (though he is not explicit on this point); his goal is to explain and justify our practice of holding ourselves responsible by illuminating that practice "in a way that enables us to understand its significance in our life."[34] The point, then, is perhaps that our practice of holding ourselves responsible has a significance in virtue of which it plays an essential role in maintaining our self-conceptions as agents, not that we should (or perhaps even could) in particular cases hold ourselves responsible *for the reason that* doing so upholds our self-conceptions as agents.

It is worth pausing here to appreciate the resourcefulness of Raz's proposal. By introducing the idea of a domain of secure competence and elaborating its role in our agency, Raz connects responsibility to what may have seemed a largely unrelated area of human concern: our vital interest in maintaining our self-conceptions as agents. This furnishes us with some useful new tools for understanding our responsibility practices, and in the next section I will put these to use in a different way.

But first I want briefly to introduce two potentially troublesome issues that Raz's account faces. I do this not in order to mount objections to the account or to raise the question of whether it succeeds on its own terms (this is beyond my scope[35]). My aim is rather to identify two questions that any account of this general kind will need to be able to answer.

We have seen Raz to argue, first, that our self-conceptions as agents are threatened when our powers of rational agency fail within our domains of secure competence; and second, that holding oneself responsible for those failures is a way of removing the threat that they pose to our self-conceptions. But both steps, on closer inspection, raise questions.

In the first place, why should we accept the strong claim that *any* failure of our powers of rational agency (aside from those due to recognized competence-defeating conditions) stands as a threat our self-conceptions as agents? Since we know that we are not infallible, it must be consistent with our self-conceptions that we are subject to a small but real probability of failure, even when we act within our domains of secure competence. As Raz acknowledges, "We are always liable to fail to control

[33] *Ibid.* p. 245.

[34] *Ibid.* p. 246.

[35] But to lay my cards on the table, I believe the two issues below, and particularly the second, could be developed into serious problems for Raz's account.

actions within our sphere of secure competence, even when no competence-defeating condition obtains."[36] But if this is so, then why should occasional realizations of that possibility of failure—*this* time, my foot slips off the brake, etc.—threaten our self-conceptions as agents at all?[37]

In the second place, how is it that holding ourselves responsible for our failures functions to remove the threat that they are supposed to pose to our self-conceptions as agents? Raz says things like, "In acknowledging responsibility for actions due to our rational powers *we are simply affirming* that they are our secure rational powers."[38] But this requires explanation; it is, at least intuitively, one thing to "affirm" the scope of our powers of rational agency, and quite a different thing to hold ourselves responsible (for instance, by blaming, reproaching, or sanctioning ourselves). A self-reproach, or a self-sanction, is not itself an affirmation of my abilities. Raz thus needs some explanation of how it is that holding ourselves responsible can accomplish the function that his account assigns it.

For present purposes, to repeat, it is not necessary to consider what resources Raz may have to address these issues; the point is rather that these are questions that an account of this kind must answer. Having them on the table will be useful in developing out of some of Raz's materials a fuller conception of taking responsibility—the task to which I now turn.

3 Taking responsibility

There is I think a vital insight at the core of Raz's account, but I propose to develop this insight in a different direction, towards a clearer conception of our practice of *taking* responsibility for *deliberate* conduct (rather than holding ourselves responsible for inadvertent conduct). By doing this I aim to put us in a better position to appreciate the distinctive and fundamental character of taking responsibility.

As we have seen, a central feature of Raz's approach is his narrow focus on the individual who is concerned with her own conduct and its implications for her self-conception as an agent. For instance, a case in which "the glass we put on the table tumbles off it . . . [and] we tend to feel annoyance, and to blame ourselves"[39] exhibits for Raz the full structure of responsibility: a threat to the agent's self-conception that is addressed by the agent's holding herself responsible through blaming herself. This tight focus on the individual agent, considered in isolation, is both fruitful and limiting.

On the positive side of the ledger, Raz's first-personal approach marks an interesting contrast with influential accounts of responsibility that concentrate their attention on the second-personal perspective—that of the offended, blaming party who holds the agent responsible.[40] Raz's emphasis on the distinctive perspective and

---

[36] Raz, *supra* note 19, p. 246.

[37] Without an answer to this question, Raz's account would have to sacrifice much of its distinctive breadth. For a related concern, focusing on the uncertain scope of Raz's proviso that we are responsible for conduct that is the result of failures of our powers of rational agency only "provided those powers were not suspended in a way affecting the action," *see* Watson, *supra* note 25, p. 408.

[38] Raz, *supra* note 19, p. 245 (emphasis added).

[39] *Ibid.* p. 245.

[40] I have in mind here again the views of Strawson and those working within the tradition he founded, including Watson, *supra* note 5; Wallace, *supra* note 5; Pamela

concerns of the agent brings welcome and useful attention to the relatively overlooked half of transaction, and is I believe the source of his main insight.[41] But on the other side of the ledger, Raz's concern with the individual agent causes him, perhaps without noticing it, to emphasize holding oneself responsible to the neglect of taking responsibility, and more generally to overlook the social or interpersonal dimension of our responsibility practices.

As a way of developing a fuller conception of taking responsibility—of what it is, what it means, and why we do it—I propose to reckon with the social dimension of our responsibility practices that is absent from Raz's picture. This will involve integrating Razian insights about the conditions of complex human agency with some basic but important platitudes concerning human psychology and sociality. In particular, it will be useful to attend to the essential role that other agents often play in our projects and activities, and the profound way in which we are often invested in others' attitudes and dispositions. Enriching the picture in this way will naturally suggest a shift in focus from holding responsible to taking responsibility, and will bring into clearer view the distinct character and significance of taking responsibility.

The starting point, then, is Raz's notion of a domain of secure competence and his plausible suggestion that complex agency of any kind requires such a domain of competence within which we are entitled to act without reflecting on the prospects for success. We can take a first step beyond Raz's picture by observing that, for human agents living in communities, the domain of secure competence also has a social dimension and plays a critical social role. To operate as an effective agent in the world, I need to be confident of my ability to place a glass on the table or to step on the brake at the appropriate moment; but to function as a full, participating member of my community, I also need to be able to *hold myself out to others* as someone who is able to do those things. When I accept a glass of red wine from my neighbor, for instance, I implicitly represent to her that I can be trusted to wield a glass securely in the normal manner, without undue risk to her upholstery. More generally, we are creatures who relate to and rely upon one another; we need not only to see ourselves as competent within certain domains, but for others to see us and acknowledge us as competent, and to treat us accordingly as potential partners in shared activities and forms of life.

These observations are meant to be obvious and uncontroversial. Yet they are absent from Raz's discussion of complex agency and domains of secure competence; by juxtaposing them against Raz's purely asocial conception, an enriched picture of our self-conceptions as agents emerges. To fill in this picture we need only further observe that to be seen as incompetent by others, and thereby as unfit to participate in the various activities and projects that make up the life of our community, is not

---

Hieronymi, "The Force and Fairness of Blame," *Philosophical Perspectives (Ethics)* 18 (2004): pp. 115-48; Stephen Darwall, *The Second-Person Standpoint: Morality, Respect, and Accountability* (Cambridge, Massachusetts and London: Harvard University Press, 2006); T.M. Scanlon, *Moral Dimensions: Permissibility, Meaning, Blame* (Cambridge, Massachusetts and London: Harvard University Press, 2008); and many others.

[41] Another writer who puts first-personal concerns and responses at the heart of her account of responsibility is Hilary Bok in *Freedom and Responsibility* (Princeton: Princeton University Press, 1998). I cannot do justice here to Bok's sophisticated account, which is of an entirely different character than either Raz's account or the proposal I will advance below.

only to our serious material disadvantage, but also threatens our sense of who we are, of what we can do, and of our own worth. Our self-esteem, our self-respect, our pride (or shame) in who we are and what we can do, our sense of what social or cooperative activities are open to us—all of these aspects of our self-conceptions as agents depend on how we understand others to assess our domains of secure competence. That is, just as our domains of secure competence play a central role in our self-conceptions as agents, our understanding of others' *perceptions* of our domains of secure competence plays a central role in our self-conceptions as agents.[42]

And we need not stop here, for there is still more to our social self-conceptions as agents—they comprise not only our understanding of how others assess our sub-personal powers of rational agency (like memory or physical coordination) and our skills and competences (like linguistic fluency or ability to drive) but also what Strawson called the quality of our wills. Normal, well-socialized agents (like us) care deeply not only about the quality of other agents' wills towards us (as Strawson observed), but also about how other agents perceive *our* qualities of will.

That is, we care greatly and intrinsically about whether others see us as well-meaning, kind, reliable, trustworthy, and so on. This concern is grounded in our basic need to participate in cooperative, trusting, caring, and otherwise meaningful relationships. Someone who was not perceived as good-willed in various ways could never be trusted, befriended, or loved in the way that adults mutually love one another, and to realize that we were in this situation would once again threaten our very sense of who we are, of what we can do, and of our own worth.[43]

Our self-conceptions as agents—our sense of what we can do, of the activities and forms of life that are open to us, of our potential roles and accomplishments, and so on—thus significantly depend on how we understand others to assess our qualities of will. We construct our practical identities, in important part, in light of how well suited we understand others to perceive us to be to participate on mutually acceptable terms in meaningful relationships and shared activities.

As before, these further steps beyond Raz's picture do not require any ambitious or controversial claims about human psychology or experience. They require only that we attend to and appreciate the significance of a set of familiar generic observations about the character of social life as most well-socialized, non-pathological individuals experience it, most of the time.

Nonetheless, integrating these observations has a significant effect. For once these pieces are in place, we are in a position to identify an important new category of threats to our self-conceptions as agents. Part of being well-socialized, we have just

---

[42] The importance of recognition of these competences by others has also been emphasized, for different reasons, in the recognition theory literature. *See*, *e.g.*, Axel Honneth, *The Struggle for Recognition: The Moral Grammar of Social Conflicts*, trans. Joel Anderson (Cambridge, Massachusetts: The MIT Press, 1995), pp. 121-30; Joel Anderson & Axel Honneth, "Autonomy, Vulnerability, Recognition, and Justice," in John Christman & Joel Anderson (eds.), *Autonomy and the Challenges to Liberalism: New Essays* (Cambridge and New York: Cambridge University Press, 2005), pp. 127–49, 130-31.

[43] *Cf.* Pamela Hieronymi's observation, "It . . . seems quite plausible that standing in relations in which the quality of one's will is recognized, both by oneself and by others, is of considerable importance. A change in what you or another person thinks about the quality of your will, in itself, changes your relations with them," *supra* note 40, 124.

said, is that our self-conceptions as agents incorporate our understanding of how various others assess our competences and qualities of will. And this means that when we come to understand that there has been some significant negative change in how others see us, this will tend to undermine and degrade our self-conceptions as agents. In this way, the discovery that someone's assessment of you has changed for the worse (or that it has never been what you hoped it to be) can constitute a threat to your self-conception as an agent.

Often, these threats concern our understanding of how others perceive our qualities of will. This will be the case, for instance, when we are discovered to have conducted ourselves in some way—for instance, in an ill-willed or vicious manner—that belies a positive view of ourselves that we had understood others to accept, such as a view of ourselves as worthy of trust, respect, or admiration. When we perceive that we will no longer be trusted, respected, or admired in the same way as before, this tends at least to some degree to undermine and degrade our self-conceptions as agents.[44]

Recall now the first question raised for Raz's account at the end of the previous section: How exactly does the relevant conduct threaten our self-conceptions as agents? I suggested there that for Raz's account to succeed it would need to be spelled out why a simple failure to perform some action within our domain of secure competence should threaten our self-conceptions as agents. But here, with respect to the social dimension of our self-conceptions as agents, the mechanism of the threat is already fully apparent. Our self-conceptions as agents include our understanding of how others assess our qualities of will; when we act poorly, we often know from the accusation and disappointment in others' eyes (or from other signs, or indeed by inference) that the relevant assessments have changed. In this situation (at least, without some effective response from us) we cannot go on as before, secure in our previously-held self-conceptions as agents. The threat is simply that we are confronted with the fact that something we had incorporated into our self-conceptions as agents is no longer the case.

One possible response to such a threat is resignation, accompanied by a painful revision of our self-conceptions as agents to reflect the diminution in our recognized eligibility to participate in shared forms of life. But this is not the only possible reaction, particularly when the source of the problem is our own poor conduct; at least sometimes, we manage instead to respond to others' accusations and disappointment in a way that mollifies and appeases, and thereby restores us in their assessment and removes the threat.

How do we manage this? The threat consists in our awareness that, because of something we did, we are no longer (for instance) trusted in the same way as before.

---

[44] Conversely, when our conduct displays our good will or virtue in a way that we perceive to impress or to provoke the gratitude or admiration of those whose perceptions matter to us, this naturally tends to reinforce or even enhance our self-conceptions as agents. Here there is no threat, but rather an opportunity to incorporate these positive perceptions into our self-conceptions by claiming credit and by registering others' positive perceptions in feelings of gratification and pride. These positive first-personal responses could also be thought of as ways of taking responsibility, though for the reasons mentioned *supra* in note 6 the focus here will remain on negative responses.

A direct and often effective way to meet this kind of threat to our self-conceptions as agents is to manifest our commitment to vindicating our trustworthiness, good will, and general social eligibility. A convincing manifestation of this kind naturally tends to reassure others of our fitness to participate in meaningful relationships. And the way in which such a commitment to vindicate our good will and social eligibility is characteristically made manifest is precisely by experiencing and expressing contrition, and by doing what is warranted to make amends and put ourselves back in the good graces of those whom our conduct has affected—that is, by *taking responsibility* for our conduct.

This brings us to the second question raised for Raz's account at the end of the previous section: How does holding ourselves responsible serve to address the threat? I suggested there that for Raz's account to succeed it would need to be spelled out how holding ourselves responsible could perform the role that Raz's account assigns it, of rebutting the threat to our self-conceptions as agents by affirming the scope of our secure competence. But here, with respect to the social dimension of our self-conceptions as agents, the way in which taking responsibility functions to address the threat is again fully apparent. Taking responsibility reassures others of our fitness to participate in meaningful relationships and activities, because it is a manifestation of the fact that we value our relationships in the right way, that we take our infractions seriously, that we are unwilling to allow others to be disadvantaged through their reliance on our good will, and that we are resolved to do better in the future. These functions unite the complex matrix of attitudes, dispositions, and responses that are characteristic of taking responsibility, including feeling contrition towards those one has affected, being disposed to apologize and make various gestures of penitence and conciliation, and being willing to repair or compensate for the harm done.

In particular, feeling contrite or sorry towards those affected by our conduct plays a crucial role, because it demonstrates that our concern for the relationship and for the effect we have had is not merely instrumental, but intrinsic; we feel badly because we value others' recognition of our good will, and the relationships that depend on this recognition, for their own sakes. By accepting our demonstrations of contrition, those whom we have affected indicate that their perceptions of our good will have been, at least to some significant extent, restored; when we receive these indications, the threat to our self-conceptions as agents is accordingly resolved. The familiar and distinctive feeling of relief that we experience when a heartfelt apology is accepted is one mark of the significance that others' perceptions of our qualities of will have for our self-conceptions as agents.

As appeared to be the case in Raz's account, I do not wish to claim that the need to meet the threat to our self-conceptions as agents necessarily furnishes a *reason* to take responsibility, or that the connection between taking responsibility and the need to meet the threat is otherwise a rational one. My goal, similar to Raz's (as I understand it), is rather to identify and illuminate our practice of taking responsibility "in a way that enables us to understand its significance in our life."[45] The point then is that our practice of taking responsibility has a significance in virtue of which it plays a role in maintaining our self-conceptions as agents, not that we should (or perhaps even could) take responsibility *for the reason that* doing so upholds our self-conceptions as agents.

[45] Raz, *supra* note 19, p. 246.

This proposal thus vindicates, via an alternative route, Raz's insight that our conduct can threaten our self-conceptions as agents and that our responsibility practices can be understood as a way of registering, responding to, and (when conditions are favorable) resolving such threats. But in structure it is quite different from Raz's account. For Raz, holding oneself responsible was conceived as an autopoietic act that reestablished the agent's self-conception in a direct, unmediated way. In this alternative proposal taking responsibility's connection to the agent's self-conception is mediated by its interpersonal significance. The advantage of this approach is that the social dimension of our self-conceptions as agents helps to explain the aptness of taking responsibility as a response to threats to our self-conceptions. If the threat consists in our awareness of diminished eligibility in others' eyes to participate in meaningful relationships and activities with them, then an intrinsic concern for what has been done to those others and the relationship, together with a readiness to conciliate them and restore ourselves in their good graces, is just the kind of response that is called for.

Moreover, the way in which, on this view, our self-conceptions as agents are bound up with our concern for others and for the relationships and activities that we participate in with them helps to explain the special, personal sense of urgency that is often a part of feeling sorry. If I hurt or offend someone, plausibly part of what makes taking responsibility make sense as a response is that I cannot tolerate being thought of as someone who holds the other and our relationship in so little regard—that's not who I am.

Before concluding, however, there is also a significant complication that is introduced by the social dimension of our self-conceptions as agents. The practice of taking responsibility just described centrally involves our concern for others' perceptions of us. But where perceptions are involved there is the possibility of error; this means that, in addition to the paradigmatic scenario in which you and I have common knowledge of the relevant facts (of the fact, for instance, that I have somehow mistreated you), there are two other scenarios to consider. In the first, I have (for instance) mistreated you but you are unaware of this; in the second, you mistakenly take me to have mistreated you when I have not. The foregoing discussion suggests that in the first case there would be no threat to my self-conception, while in the second case there would be a threat; and this might seem to entail that it would be inapt for me to take responsibility in the first case but would be apt to do so in the second. On first blush, this seems both psychologically and normatively questionable. Moreover, it seems to suggest a significant divergence between our practice of taking responsibility and our considered judgments of when an agent actually is responsible.

On fuller consideration, I think these apparent difficulties are less serious than they may at first seem, though to see this will require further enriching the picture. Start with the first scenario, in which I have mistreated you but you are unaware of this. Here your perceptions of me will not have changed, and thus no revision of my self-conception as an agent is called for (at least with respect to those perceptions). And yet, having mistreated you am I not likely to feel, not merely regret, but contrition, including a desire to come clean, to apologize, and to make it up to you?

Of course, I am; but this can be made sense of in terms of a few further familiar facts. A wide swath of our emotional lives consists in our attitudes about what various other people *would* think (and in particular, would think of us), if they knew what we know. The fact that an admired grandparent would approve of my career or accomplishments can be the basis for profound gratification and pride, even if the

grandparent is long dead; entire life projects can be motivated in this way. Similarly, if I would be deeply ashamed for you to learn of something I have done, I may be uncomfortable in your presence and unable to meet your gaze, even if I have taken measures that guarantee you will never learn of it. This is because being in your presence makes salient to me what you *would* think if you knew, and this triggers intimations of the shame that I would feel—intimations that can be so strong and vivid as to be virtually indistinguishable from the real thing.

This phenomenon generalizes, and explains why it is not ordinarily enough merely to be seen by others as (for instance) trustworthy if we know privately that we are not—that is, that relevant others would not so see us if they knew everything that we know. Maintaining such deceptions is ordinarily cognitively and emotionally draining, and also, at least with respect to our self-conceptions as agents, largely ineffectual. This is because our self-conceptions as agents tend to be sensitive, not only to what people actually think of us, but also to what they would think of us if they knew what we know. And an important upshot of this is that ordinarily the most effective way to feel secure in others' trust is to convince oneself that one is in fact worthy of it. The knowledge that one is less trustworthy than others believe is thus often enough to subvert one's self-conception as an agent, in the extended sense being developed here, even if others have not discovered this fact.

It follows that my mistreatment of you can be the basis for feelings of contrition towards you, through my understanding of what you would think if you knew what I had done, even if I am confident that I can prevent you from ever knowing. And this is, moreover, both a basic psychological fact about us and a normatively attractive one that we tend to endorse on reflection and to value and cultivate in ourselves and one another.

Consider now the second scenario, in which you mistakenly take me to have mistreated you when I have not. Here there is (let's say) a change in your perceptions that threatens to disrupt my self-conception as someone whom you trust; and this would seem to render it apt for me to take responsibility—to feel contrition, to be disposed to apologize, and so on, in order to be restored in your good graces and resolve the threat. But doesn't this seem odd, or even perverse?

In a sense, admittedly, it is—but not, I think, in a way that should be taken to undermine the picture we have developed. First, we should not exaggerate the extent of the oddness; second, we should not forget that many aspects of our emotional lives can appear odd or perverse from the sometimes austere perspective of moral theory.

Regarding the extent of the oddness, it is important to take into account the fact that any pressure to take responsibility in this scenario is liable be swamped or undercut by other factors. After all, where the threat to one's self-conception as an agent consist in another's error, it can often be deflected simply by correcting the error, for instance with an explanation or by avowing or presenting evidence of innocence. Taking responsibility is, moreover, frequently emotionally, and indeed materially, onerous; so it is easy to see why it would not present itself as the natural response in circumstances where pleading innocence may be an effective alternative. It is also significant that while others' perceptions of our good will are important for our self-conceptions as agents, they are not the only important thing; taking responsibility as a response to another's false perceptions is something we are often strongly disinclined to do on a number of grounds, including pride, a sense of integrity, and a commitment to basic fairness. These considerations are further bolstered by the fact that taking

responsibility, even when it is successful in restoring us in others' good graces, is often humbling, and sometimes humiliating. To be humbled for what we have done is one thing; but to be humbled for what we have not in fact done often feels to us intolerable. And this, again, is a pattern of response that we tend to endorse on reflection and to encourage and cultivate.

We should not, then, overestimate the pressure that false perceptions can put on us, in versions of this second scenario, to take responsibility. Yet at the same time, we should not indulge in the rationalistic illusion that such pressure does not exist. The blame or disapproval of those whose views matter to us can touch us very deeply. Even where we have the strength of conviction, or other emotional resources, sufficient to refuse responsibility and resist contrition, the pressure exerted by others' blame is frequently redirected and appears in other forms, such as a sense of grievance or indignation ("How *could* she think me capable of that?"). A different kind of redirection occurs when the sense of contrition is, in a simultaneously conciliatory and face-saving maneuver, attached to an object other than the perceived fault ("I'm sorry that this misunderstanding has arisen between us"[46]).

Conversely, the pressure of blame will be most likely to provoke a direct response in the form of contrition for something the agent did not do, in versions of the scenario in which simply pleading innocence is not a viable alternative (for example, because the agent lacks the credibility or standing to do so effectively), and where the agent lacks the maturity, pride, integrity, or sense of self that otherwise might provide a counterbalance. This means that the tendency to take responsibility and to feel contrition in the face of false perceptions will distribute itself in familiar ways around social hierarchies and other disparities of power, and will accordingly be subject to exploitation for purposes manipulation or domination. For these reasons, it seems best to discourage the tendency to take responsibility and to feel contrition in these circumstances—as we might say, inappropriately or irrationally. We might even prefer to withhold the labels, and to say that in this scenario these responses would not count as real contrition, or as the real taking of responsibility, but only as deviant or degenerate cases. That is fine, and perhaps salutary, but need not prevent us from appreciating the structural and other similarities between these inappropriate or deviant responses and more paradigmatic cases of taking responsibility.[47]

Finally, in light of this how should we understand the relation between our practice of taking responsibility and what it is to *be* responsible more generally? As I noted at the outset, the nature and conditions of responsibility are not my topic; my more limited aim has been to advance our understanding of certain responses and practices that are associated with responsibility. Nevertheless, we would expect to find some close and systematic connection between our practice of taking responsibility and what it is to be responsible. I have offered some reasons to think that the divergences between them may not be as wide as a first glance at scenarios involving false perceptions might suggest, but those divergences cannot be made to disappear altogether; according to the picture that we have developed we would not expect, even in a well-

[46] Here the useful ambiguity of "sorry" as a possible expression of either contrition or impersonal regret often plays an important role.

[47] For a related point about the appropriateness of emotional responses, *see* Justin D'Arms & Daniel Jacobson, "The Moralistic Fallacy: On the 'Appropriateness' of Emotions," *Philosophy and Phenomenological Research* LXI (2000): pp. 65-90.

socialized and self-possessed agent, that contrition and responsibility will always go hand in hand.

What this means is that if we aim, in a Strawsonian spirit, to understand what it is to be responsible in terms of our practice of taking responsibility (and to repeat, this has not been my aim here), this understanding will have to involve some degree of idealization. We cannot simply say, for instance, that for an agent to be responsible *just is* for it to be apt for her to take responsibility. Any plausible account along these lines will have to incorporate at least some qualifications and conditions concerning (no doubt among other things) the knowledge and states of mind of the relevant parties. But this should be expected; our actual practice of taking responsibility (like our practice of holding responsible) is a very human and therefore untidy institution, whereas our idea of what it is to be responsible is a moralized abstraction. It is no surprise if the connection between these two things, however essential, should turn out not to be entirely simple and direct.

4 Conclusion

I have sought to identify a connection between our sociality and our self-conceptions as agents, and to show how this connection can help to make sense of the unduly neglected phenomenon of taking responsibility. A key inspiration has been Raz's insight that our own conduct can threaten our self-conceptions as agents and that our responsibility practices can be understood as a way of responding to such threats, but I have developed this idea in a different way, focusing on the manner in which our self-conceptions make reference to others' attitudes towards us.

This vindicates and expands upon the picture sketched at the outset, according to which *taking responsibility* has a fundamentally different character from *holding responsible,* including holding oneself responsible. To recap: Holding oneself responsible is similar to holding another person responsible; it consists in attitudes and responses that you take towards yourself *as agent*—blame, reproach, sanction, and so on. Taking responsibility, on the other hand, functions as a response to threats of a certain kind to your social self-conception, and consistent with this orientation it consists of attitudes and responses you take towards the other *as someone affected by your conduct* and *as a party to the relationship*. Taking responsibility is thus complementary to but distinct from holding responsible.

If this is right then taking responsibility turns out to be independently important and theoretically interesting, both in itself and, at least arguably, for the sake of understanding what it is to *be* responsible. Return to the case in which you cancel a meeting, disappointing your friend. It may be that your friend is merely disappointed rather than inclined to blame you or otherwise hold you responsible for this, and even that it would not be appropriate for her to do so under the circumstances (you had excellent reason!); but even so, perhaps her disappointment lingers, casting a shadow over your relationship. Even if there are no grounds for blame, we have no difficulty whatever in grasping the notion that you are nonetheless responsible for cancelling the meeting—not merely in the sense that canceling was something that you in fact did or that it was an expression of your agency or that it was attributable to you, but also in the further sense that in light of what you've done it makes sense to feel, respond, and act in ways that are distinctively associated with taking responsibility—including for instance apologizing or making amends ('making it up to her'). This suggests that to

focus solely on holding responsible is to miss an important dimension of our commonsense understanding of responsibility.

Moreover, and finally, it begins to seem doubtful whether we can even fully understand what it means to hold someone responsible without also taking proper account of taking responsibility (and vice versa). After all, everyday experience shows that the two patterns of response are not merely complementary but interactive and interdependent. The blaming attitude by which I hold you responsible can, for instance, be appeased and mollified by your demonstration of contrition; indeed, this may be the only way of truly satisfying me (rather than, say, merely distracting me, playing on my sympathies, or wearing me down). This is obviously more than just a curious regularity; it is a matter of some kind of responsiveness and fit. The tightness and apparent naturalness of these connections between holding responsible and taking responsibility suggest that these patterns of response are essentially embedded within, and to a significant extent derive their meanings and structures from, a wider emotional economy of interpersonal relations—an economy that is animated in large part by basic concerns that we share as active, social creatures, including our concern for the various meaningful connections that we can form with others, and the vital significance of these connections for our own agency. In this light, our responsibility practices are revealed as likely to be grounded in a more complex matrix of concerns and responses than has often been supposed.

Acknowledgements: For helpful conversations and comments on various versions of this material I am indebted to Alex Dietz, Erik Encarnacion, Pamela Hieronymi, Todd Jones, Greg Keating, Andrei Marmor, Abelard Podgorski, Jon Quong, Alex Sarch, Mark Schroeder, Beth Snyder, Gary Watson, and Aness Webster, as well as participants in the 2014-15 USC Dissertation Seminar and an audience at the 2015 SoCal Philosophy Conference. Special thanks are due to two anonymous reviewers, including a reviewer for this journal who provided exceptionally helpful and constructive comments. Work on this paper has been supported in part by a University of Southern California Provost's Ph.D. Fellowship and a Ralph and Francine Flewelling Graduate Fellowship.

References

Joel Anderson & Axel Honneth, "Autonomy, Vulnerability, Recognition, and Justice," in John Christman & Joel Anderson (eds.), *Autonomy and the Challenges to Liberalism: New Essays* (Cambridge and New York: Cambridge University Press, 2005), pp. 127–49.

Christopher Bennett, *The Apology Ritual: A Philosophical Theory of Punishment* (Cambridge and New York: Cambridge University Press, 2008).

Hilary Bok, *Freedom and Responsibility* (Princeton: Princeton University Press, 1998).

Justin D'Arms & Daniel Jacobson, "The Moralistic Fallacy: On the 'Appropriateness' of Emotions," *Philosophy and Phenomenological Research* LXI (2000): pp. 65-90.

Stephen Darwall, *The Second-Person Standpoint: Morality, Respect, and Accountability* (Cambridge, Massachusetts and London: Harvard University Press, 2006).

John Deigh, "Love, Guilt, and the Sense of Justice," in *The Sources of Moral Agency* (Cambridge and New York: Cambridge University Press, 1996), pp. 39-64.

R.A. Duff, *Punishment, Communication, and Community* (Oxford and New York: Oxford University Press, 2001).

David Enoch, "Being Responsible, Taking Responsibility, and Penumbral Agency," in Ulrike Heuer & Gerald Lang (eds.), *Luck, Value, and Commitment: Themes from the Ethics of Bernard Williams* (Oxford and New York: Oxford University Press, 2012), pp. 95-132.

David Enoch, "Tort Liability and Taking Responsibility," in John Oberdiek (ed.), *Philosophical Foundations of the Law of Torts* (Oxford and New York: Oxford University Press, 2014), pp. 250-71.

David Enoch & Andrei Marmor, "The Case Against Moral Luck," *Law and Philosophy* 26 (2007): pp. 405-36.

Pamela Hieronymi, "The Force and Fairness of Blame," *Philosophical Perspectives (Ethics)* 18 (2004): pp. 115-48.

Axel Honneth, *The Struggle for Recognition: The Moral Grammar of Social Conflicts*, trans. Joel Anderson (Cambridge, Massachusetts: The MIT Press, 1995).

Coleen Macnamara, "Holding Others Responsible," *Philosophical Studies* 152 (2011): pp. 81-102.

Hannah Maslen, *Remorse, Penal Theory and Sentencing* (Oxford and Portland, Oregon: Hart Publishing, 2015).

Linda Radzik, *Making Amends: Atonement in Morality, Law, and Politics* (Oxford and New York: Oxford University Press, 2009).

Joseph Raz, *From Normativity to Responsibility* (Oxford and New York: Oxford University Press, 2011).

T.M. Scanlon, *Moral Dimensions: Permissibility, Meaning, Blame* (Cambridge, Massachusetts and London: Harvard University Press, 2008).

Angela M. Smith, "Responsibility for Attitudes: Activity and Passivity in Mental Life," *Ethics* 115 (2) (2005): pp. 236-71.

Nick Smith, *I Was Wrong: The Meanings of Apologies* (Cambridge and New York: Cambridge University Press, 2008).

Peter Strawson, "Freedom and Resentment," in Gary Watson (ed.), *Free Will* (Oxford and New York: Oxford University Press, 2003, 2nd ed.), pp. 72-93.

John Tasioulas, "Punishment and Repentance," *Philosophy* 81(2) (2006): pp. 279-322.

Gabriele Taylor, *Pride, Shame, and Guilt: Emotions of Self-Assessment* (Oxford and New York: Clarendon Press, Oxford University Press, 1985).

Margaret Urban Walker, *Moral Repair: Reconstructing Moral Relations After Wrongdoing* (Cambridge and New York: Cambridge University Press, 2006).

R. Jay Wallace, *Responsibility and the Moral Sentiments* (Cambridge, Massachusetts and London: Harvard University Press, 1994).

Gary Watson, "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme," in *Agency and Answerability* (Oxford and New York: Oxford University Press, 2004), pp. 219-59.

Gary Watson, "Peter Strawson on Responsibility and Sociality," in David Shoemaker & Neal A. Tognazzini (eds), *Oxford Studies in Agency and Responsibility, Volume 2: "Freedom and Resentment" at 50* (Oxford and New York: Oxford University Press, 2014), pp. 15-32.

Gary Watson, "Raz on Responsibility," *Criminal Law and Philosophy* 10 (2016): pp. 395-409.

Bernard Williams, "Moral Luck," in *Moral Luck: Philosophical Papers 1973-1980* (Cambridge and New York: Cambridge University Press, 1982), pp. 20-39.