Ideas y Valores: Revista Colombiana de Filosofía (forthcoming)

Reasons and Real Selves

Manuel Vargas
University of San Francisco

Most accounts of responsibility begin from either of two prominent points of departure: the idea that an agent must have some characterological or expressive connection to the action, or alternately, the idea that an agent must be in some sense responsive to reasons.[1] Indeed, we might even understand much of the past couple of decades of philosophical work on moral responsibility as concerned with investigating which of these two approaches offers the most viable account of moral responsibility. Here, I wish to revisit an idea basic to all of this work. That is, I consider whether there is even a fundamental distinction between these approaches. I will argue that the relationship between these two approaches to moral responsibility is much more complicated than is ordinarily assumed. I shall argue that there are reasons to think that one of these views may ultimately collapse into the other, and if not, that there is nevertheless reason to think one of these views has misidentified the features of agency relevant to moral responsibility. The view that follows is one that we might call *the primacy of reasons*. In the second half of the article I consider whether recent experimental work speaks in favor of the alternative to the primacy of reasons. Its proponents argue that it does. I argue that it does not.

1. Real Selves and Reasons: Some initial considerations

One inspiration for those accounts of responsibility that emphasize a characterological or expressive connection between agent and action is the idea that it can only make sense to hold someone responsible if the action in some way expresses a deep fact about the particular agent. Contemporary versions have variously emphasized that the agent needs to "identify" with the motives that lead to the act, or the act has to be expressive of a "Real Self" or the agent's values, or the action has to be an expression of the regard in which the agent holds others.[2] Following the customary parlance given to us by Susan Wolf, I will call such accounts Real Self views, or RS views.[3] The label is imperfect, for it is not obvious that all accounts that appeal to a condition of identification or self-expression need be committed to the existence of a "real self" in any substantive way. Nevertheless it is a serviceable misnomer because it emphasizes the idea of some special or privileged subset of psychological states in relation to which the agent's actions must stand for there to be responsibility.

If RS accounts emphasize that the mark of responsible agency is the presence of psychological structures that, roughly, express some privileged view of the agent's, the mark of its alternative —Reasons accounts— is the presence of a particular power to respond to the world. On these latter accounts, the agential contribution to responsibility is a power to respond to the reasons that arise from the world or the agent's psychology's interaction with the world. On this account, what makes responsible agency distinctive is that the agent's response to the world is structured by reasons in a particular way. It is not the projection of the agent's identity or convictions that makes action responsible but rather how the agent's actions express (or don't) due sensitivity to reasons.[4]

Characterized in this way, the difference between RS accounts and Reasons accounts may seem extraordinarily thin. One might wonder whether the manner in which one responds to reasons is just a way of expressing one's character, commitments, or values. And, one might suspect that one's character, commitments, and values say something about what the agent regards as reasons-giving. If so, then even if RS views and Reasons accounts can claim to have different points of theoretical departure, those departure points are surely not far apart. At the very least it suggests that the subject matter of theories of moral responsibility is not so radically bifurcated so as to suggest that there are distinct phenomena that are mistakenly given the single label of 'moral responsibility'.

Whatever the similarity of starting point, RS approaches face some distinctive worries. Consider, for example, a paradigmatic RS account—Harry Frankfurt's, as presented in the 1971 article "Freedom of the Will and the Concept of a Person."[5] On Frankfurt's account, an agent is responsible for some action if and only if at the time of action the agent had a particular second order desire— i.e., a desire that the motivating first order desire be effective in action. Notice that the higher order desire need not be causally efficacious itself— it could be "along for the ride," so to speak, and its presence or absence might play no causal role in whether the agent acts on some particular first order desire. So, on Frankfurt's account, a willing addict is morally responsible for a decision to take his drug of choice even if the higher order desire that one act on the drug-taking desire plays no causal or explanatory role in the taking of the drug.

Frankfurt's account and its subsequent developments have been construed in different ways —as, for example, a picture of autonomy, of free will, of responsible agency, of "strong agency" and so on.[6] Construed as an account of the kind of agency required for moral responsibility, however, the picture in "Freedom of the Will" provides at least three reasons for consternation. First, on Frankfurt's account, all that matters for securing responsibility is the presence of the requisite psychological structure, regardless of its origin. This entails some startlingly counterintuitive possibilities. For example, an agent that has an alien set of values transplanted by coercive indoctrination, brainwashing, or (currently science fictional) neurological implantation would, it seems, count as straightforwardly responsible for any action subsequent to the implantation. Second, an insufficiently knowledgeable, or a systematically delusional agent, is hardly a model of responsibility, no matter how self-identified.[7] Yet, on Frankfurt's account it seems that we must say that such an agent is a responsible agent. A natural way to respond to such worries is to appeal to the rationality of the agent's beliefs, or to a connection between agents, norms, and the structure of the world. But if we supplement the account in this way, then it looks less distinctive as an alternative to reasons accounts. Thirdly, the account is silent on the matter of why, precisely, it is that second order desires are the sort of thing that provide a basis for moral responsibility.[8] A second order desire is still a desire, and the fact of it being of the second order does not seem to, by itself, constitute any reason to regard it as expressing where the agent stands. One way of understanding the criticism is that it is unclear why the fact of where some agent stands, were it tractable in terms of hierarchies of desires, should be the kind of thing in virtue of which moral praise and blame make sense.

I wish to focus on this latter criticism, that we need some account of why those psychological elements identified by a RS account are sufficient for grounding the appropriateness of praise and blame. One way to appreciate the force of the worry is to consider an appeal to psychological states that are manifestly irrelevant to grounding moral praise and blame. For example, if someone were to argue that it was hierarchies of jealousy, or hierarchies of beliefs, or hierarchies of hunger that determined the appropriateness of moral responsibility, we would surely demand an explanation of why such things are at all relevant to moral responsibility. In the case of desires, the idea that they have some connection to warranting praise and blame is an old one. Nevertheless, we can and should ask *why* hierarchies of desires should be the sort of things that warrant praise and blame.

There are a number of things the Real Self theorist might say in the face of this challenge.[9] For example, perhaps the reason why higher order capacities are significant for moral responsibility is precisely because they reflect some further fact about the agent, and in virtue of that further fact, praise and blame come to make sense. Frankfurt suggests something very much like this in the context of considering whether creatures other than humans might count as having higher order desires. He writes "No animal other than man, however, appears to have the capacity for reflective self-evaluation that is manifested in the formation of second-order desires" (12). If I understand Frankfurt rightly, his claim is that there is a comparatively unusual capacity required before one can have second-order desires, something he calls "the capacity for reflective self-evaluation." It is this capacity that sets humans apart from other animals, and in virtue of which we come to be able to have second-order desires. Higher order desires are a kind of proof for its existence, for one could not have such desires without a capacity for reflective self-evaluation. So, perhaps, the thought is that those capacities are part of what makes the presence or absence of hierarchies of desire relevant for moral responsibility.

The existence of this enabling capacity raises some puzzles about Frankfurt's account. In particular: what is doing the explanatory or normative work in the account? If higher order evaluations are really products of some more basic feature, why not look to that distinctive capacity as the locus of freedom, personhood, and moral responsibility? Indeed, what seems to give those higher order desires any force or relevance at all for the matter of responsibility is that they are the products of reflective self-evaluation. For example, if they were simply brute desires, or products of unmediated instinct, it would be difficult to see how they could support the distinction Frankfurt is looking for, one where on one side we have unremarkable animals, and on the other side we have agents capable of personhood, freedom, and moral responsibility. What makes second-order desires special seems to be precisely that they are the products of reflective self-evaluation. So, perhaps what Frankfurt *should* have said is that it is not second-order desires, per se, that matter for distinguishing responsible agents from nonresponsible agents. Responsible agents are, in some way, a byproduct of a more fundamentally important capacity, and it is something about this underlying capacity that makes sense of the appropriateness of praising and blaming.

One consequence of replying in this way is that Frankfurt's account threatens to collapse into a *de facto* Reasons account. It is difficult to see how the capacity for self-reflection is not just self-directed rational assessment. Frankfurt's "reflective self-evaluation" seems to be a self-aware, self-directed form of those capacities emphasized by Reasons accounts: i.e., reflective self-control, or the capacity to recognize and appropriately respond to reasons.[10] To be sure, he seems to have in mind a particular subset, or perhaps a particular application of those abilities—namely, those tied to self-awareness. Still, ultimately we are left with an appeal to a species of rational power. If so, then we have come to the startling conclusion that Frankfurt's account is really committed to a species of the Reasons approach.

The foregoing suggests that the paradigmatic RS account is not itself a genuine, distinctive option in the way ordinarily regarded in the literature.[11] In turn this might suggest a view we can call *the primacy of reasons*. On this view, our rational capacities are central to moral responsibility, and purportedly alternative accounts will, on closer inspection, either smuggle in a commitment to rational capacities or prove to be inadequate. In the face of such a view, one could rightly object that even if one accepts that Frankfurt's account is vulnerable to concerns about its force deriving from the role of rational powers, this need not be true of any and every RS account. It would require a good deal more discussion than I have offered to show that other or all RS accounts ultimately bottom out into a story of rational powers. Fair enough. Still, we might take the present reflections to generate a challenge to extant RS accounts: is there any reason to think that the psychological features highlighted by one's preferred RS account have some special status, apart from their genesis in the rational faculties of agents? Put differently, what RS theorists need are two interconnected things: (1) an account of why the psychological structures they identify are the kinds of things in virtue of which agents can be responsible, and (2) an explanation of why the normative relevance of those structures is not ultimately parasitic on, or reducible to the exercise of the capacities that constitute the heart of Reasons accounts.

2. A new argument for RS theories?

Going forward, I will assume that RS accounts face the two-pronged challenge mentioned above. Now I wish to consider one way in which the proponents of RS accounts might reply. In its basic elements, the reply is this: the reason those psychological structures appealed to on a RS view count as the features in virtue of which agents are responsible is that those structures are the focus of our existing judgments of responsibility. Inasmuch as a theory of responsibility is properly beholden to our ordinary judgments about cases, we answer the "why these structures?" question by appeal to their centrality in our responsibility assessments. So, even if these structures are parasitic on reasoning capacities in some fundamental way, it is those higher-level psychological structures to which we are responding in our responsibility assessments, and it is the presence of these specific structures (and possibly, the absence of specific structures or properties) that constitute one's being a responsible agent. On this account, the gap between those psychological structures and the warrant for praise and blame is bridged by our basic epistemology of moral responsibility.

One virtue of this reply is that it permits the RS theorist to concede a kind of dependence on underlying rational capacities, without thereby surrendering the need for a distinctively RS account of moral responsibility. However, for this strategy to succeed it needs some warrant to motivate its central claim that RS views are uniquely good at capturing the phenomena of ordinary judgments of responsibility. Fortunately for the RS theorist, there appears to be some evidence of just this sort.

In a recent discussion of the relevance of experimental data for moral theorizing, Doris and Stich have pointed to a series of provocative experiments conducted by Woolfolk, Doris, and Darley.[12] What these experiments seem to show is that attributions of responsibility tend to track an agent's identification with the action; identification or its absence is the most salient trigger of our assessments of responsibility. If that is correct, then this is exactly the sort of evidence the RS theorist might hope to find: evidence for a tight conceptual link between RS-favored psychological structures and the warrant for praise and blame.

Woolfolk et al. have subjects consider a scenario in which they are told about two couples that are friends, returning from a vacation together. One of the members of this group of four adults, Bill, has learned that his wife (Susan) and his best friend (Frank) have been involved in an illicit love affair with each other. The subjects are told that Bill has just discovered proof. The subjects are then given one of several different versions of the case. In the low identification version of the case Bill decides that he is going to confront Susan and Frank, but that he has resolved not to stand in their way if they want to be together. In the high-identification version of the case, Bill decides that he will kill Frank. The philosophically interesting results emerge in the high identification case. Subjects in the high identification version of the case are told that before Bill does anything, hijackers take over the plane and things eventually get to a situation where Bill is ordered by the hijackers to shoot Frank, and he does so. What Woolfolk et al. discovered was that subjects are more willing to judge high-identification Bill as more responsible, more appropriately blamed, and more properly subject to guilt than low-identification Bill. Even more remarkably, this was so even in scenarios where the hijackers were described as having additionally administered to Bill a "compliance drug" that forced him to behave exactly as they ordered. That is, even in the presence of multiple overdetermining elements to Bill's action, subjects were more willing to hold high-identification Bill responsible, as compared to low-identification Bill. So, what Woolfolk et al. seemed to have found was that ascriptions of responsbility very tightly track identification.[13]

In light of results such as these, the RS theorist might have some reason to claim that RS theories are uniquely well-suited to capturing distinctive phenomena of the sort manifested in the Woolfolk et. al. results. (Indeed, one could even think that not only do these results favor a specifically identificationist RS account, they even suggest—as Frankfurt himself famously argued—that alternative possibilities are no requirement on moral responsibility.[14]) So, one might think we have an answer to the challenge facing RS views. The evidence for our tracking identification in responsibility ascriptions seems to support the idea that identification is central to our concept and practices of moral responsibility.

3. Against the new argument for RS theories

The Woolfolk et al. results are provocative, but less than the RS theorist needs. First, it is far from clear that empirical data alone will be sufficient to demonstrate that RS accounts can explain the special status of those psychological features they identify, apart from their relationship to rational capacities. That is, even if there are some phenomena that RS theories are particularly well-suited to explain—let us suppose the Woolfolk et al. capture such phenomena—these considerations have to be balanced against the costs of accepting the theory, especially given the costs and benefits of alternative accounts. (Here, recall the aforementioned worries regarding manipulation and sanity.) So, the most we can expect from data of this sort is support for *one* premise in a more complicated argument for RS views.

There is a second, and more powerful reason to be doubtful about the overall utility of these examples. To put it simply: you don't need an RS theory to account for these results. To see why, think about the general issue of how we become responsible for what flows from our habits and character. A very natural way to accommodate the idea that we are responsible for actions deriving from character and habit is to think that our choices shape us, and that in turn, we are shaped by those features of our character that are built up out of individual choices. On this picture, as we make choices they slowly come to form settled habits of character.[15] Sometimes this operates on the basis of habituation. In other cases, it might arise as a consequence of settling on an explicit, self-governing policy that filters the agent's downstream deliberative options.[16] If I have a policy of starting the coffee pot immediately after getting out of bed, this policy will typically have the result of filtering out other deliberative options when I get out of bed (e.g., checking email, reading the latest news, firing up the waffle iron, etc.). Whether by habit or self-governing policy or both, prior choices can permit us to extend our powers of agency into the future in comparatively stable and reliable ways.

Considerations such as these have given rise to widespread acceptance of what can be called *a tracing theory* of moral responsibility.[17] On this picture, one way we can be responsible for what we do is by being responsible for who we are. This capacity is important, as much of what we do is a product of habits, policies, and character traits. It is by being responsible for the formation of these habits, policies, and character traits that we come to be responsible for much of what we do. That is, we can trace our responsibility for actions that derive from habits, policies, and character traits back to our antecedent choices that led to those aspects of ourselves.

Tracing is an important part of the repertoire of most theories of responsibility. Tracing helps to explain away many of the cases that might otherwise appear to be accommodated only by an RS account. Tracing does this by permitting us to say that for any putative instance of responsibility, responsibility need not be accounted for by appeal to the presence of (for example) rational capacities at the time of action. Instead, all that is needed is some prior decision, character trait, habit, or policy that itself constitutes responsible choice (where this includes possession of some suitable knowledge), under conditions where those things were arrived at through the operations of the requisite agential features. So, suppose Kevin has the deplorable policy of insulting any student who comes to speak to him during office hours. And, suppose that Kevin is no longer reflective at all about this practice, and not sensitive to moral considerations that weigh against it. However, when Kevin formed this policy he was alive to those considerations and simply decided to dismiss them—perhaps even welcoming their deterrent effect on students visiting him during office hours. Now, though, when Kevin's students arrive to office hours, he habitually says (with a loud chuckle): "What stupid question are you too dumb to answer on your own?" On a tracing theory, the most natural thing to say about Kevin is that he is responsible for insulting his student. After all, he was carrying out a policy that he formed freely and responsibly (e.g., on a Reasons account, under conditions of rational self-governance). That his later deployment of that policy was unreflective and automatic is immaterial given the presence of that prior anchor in suitable features of agency.[18] Similarly, a drunk driver does not get off the moral hook simply because at the time he hit someone with his car he was especially intoxicated, and thus not responsive to reasons. In such cases, we look back to earlier decisions to, for example, begin drinking when there was reason to think one might come to drive, or in adopting habits of excessive drinking, or in deciding against being cautious about the risks of drinking, and so on.

Once we recognize the possibility of tracing, it is difficult to see how examples of the sort generated by Woolfolk et al. require an RS view. Opponents of RS views will simply insist that Bill's responsibility for his killing Frank is grounded in his (free) decision to kill Frank, prior to the actions of the hijackers. While Bill might not have envisioned the particular details of how we was going to kill Frank, his deciding to do so is a sufficient anchor for tracing responsibility. As long as there is no reason to suppose the prior decision violated one's (non-RS) conditions of responsible agency, then there is no reason to rule out this sort of tracing. That the hijackers coerced Bill might involve some diminution of responsibility—which is, anyway, consistent with the responses Woolfolk, et al. received. However, such concern does not mean that Bill cannot be held morally responsible for pulling the trigger.

So, a critic of RS theories is unlikely to be moved by the Woolfolk et. al. evidence. However, the proponent of an RS theory will surely object that there is a crucial element of the results that have not yet been addressed: where there is more identification there is more willingness to ascribe responsibility. Indeed, one might think, this is the most important result arising from those experiments. So, the RS proponent might say, even if critics can explain why people might think Bill is responsible in cases where there is a compliance drug present, the data still supports the idea that what is central to our ascriptions of responsibility is identification.

However, there is a natural reply to be made to this point as well. While it is true that the data provide something of an initial warrant for thinking that identification is central to how we ascribe and think about the requirements for moral responsibility, this is also consistent with thinking that identification matters to us only *evidentially*. That is, our tracking whether an agent identifies with some outcome or act is a piece of evidence for some more metaphysically or normatively salient property to which identification points. So, for example, suppose we had the view that responsibility depends on, roughly, (1) whether an agent is capable of rational self-governance in that particular context and (2) whether the agent has done something morally wrong. On this view, we ordinarily have good reason to track whether an agent identifies with his or her action in a given context. Whether an agent identifies with an act counts as a good piece of evidence *for thinking the agent has the relevant rational capacities in that context*. Why think that? Well, one might think it for exactly the sort of reason suggested by Frankfurt in "Freedom of the Will and the Concept of a Person": identification *strongly* suggests —even ultimately requires— the presence of rational self-governance. Where there is identification there is rational, reflective agency. Note, moreover, that this is a perfectly general point, one that does not necessarily require a Reasons view. For example, suppose you thought that the central agential feature that is crucial to moral responsibility is the presence or absence of ill will.[19] On such a view, identification will plausibly be important to our epistemology of responsibility. However, its importance is derivative. It is a byproduct of our inability to directly access what we are really interested in, whether it is ill will, rational capacities, or something else.

So, it seems, the Woolfolk et al. data do not settle the matter or even obviously favor the RS theorist. Consequently, RS theorists have not yet identified a special conceptual connection between, on the one hand, praise and blame and on the other hand, those psychological structures implicated by RS views.[20] Minimally, what is required is a different set of experimental results, results whose experimental model controls for the possibility that identification (or some other RS property) has only an evidential role to play. Until we see such an experiment and the attendant results, it seems that the RS theorist cannot appeal to experimental data for forging a link between the theory's preferred psychological structures and praise and blame.

4. Is the best defense is a good offense?

Thus far, I have argued that the familiar distinction between RS and Reasons approaches to moral responsibility is less clear than one might think. In particular, I have argued that RS views are under pressure to show that there is some reason to think that the psychological features highlighted by one's preferred RS account have a special status, apart from being evidence of the rational faculties of agents. If they cannot show this, then it suggests that RS theorists fail to have a distinctive approach to accounting for moral responsibility, and more importantly, that the focus on a "real self" constitutes a misidentification of the features of agency in virtue of which moral responsibility obtains. I then considered a line of reply that makes use of recent empirical data suggesting that ordinary attributions of responsibility tightly track identification. In reply, I noted that appeals to tracing and the evidential role of identification permit non-RS accounts to explain away the experimental evidence. The appeal to tracing, though, is important. Without tracing, the principal alternative to RS accounts —Reasons accounts—do not obviously have the resources to explain away the persistence of our responsibility attributions under conditions where agents do not seem to be actively exercising rational capacities.

It is on this issue where RS theorists might plausibly go on the offensive. Although tracing is common in the literature of responsible agency, it has been recently argued that these accounts are plagued by an under-appreciated difficulty. The difficulty is this: in many circumstances the anchoring traits, habits, or policies are adopted under conditions in which the agent has poor epistemic access to the consequences that flow from having adopted that trait, habit, or policy.[21] It is perhaps a truism that I cannot be held responsible for some outcome unless it was reasonably foreseeable—except where my lack of foresight is itself something for which I am responsible. In the context of tracing theories, the worry is this: in a wide range of cases, the aspects of our self, character, or policy which provide the basis for many of our actions were acquired in circumstances under which we could not foresee the implications for our future actions of our acquiring them. Or, to put it somewhat differently, the anchors for our responsibility traces cannot secure responsibility when the downstream effect was not reasonably foreseeable at the time of the anchoring decision. Indeed, the more remote—temporally or recognitionally—the context of action is from the context of the acquisition of the trait, habit, or policy, the more significant we should expect the epistemic defect to become. Many of the characteristics I inculcated in myself in junior high school were doubtlessly acquired under conditions when I would or could not know about their consequences in my more mature adult life.

How ubiquitous this problem is remains an open question. As a problem for theories of responsibility, it depends in part on the frequency with which the theory relies on tracing. Accounts that hold that we have free will somewhat infrequently will face a version of this problem to a greater extent than theories that require little or no tracing, or whose tracing does not typically involve significant temporal extendedness. It is on this point, however, that the thin edge of the RS wedge might be inserted. Earlier, I noted that RS theorists could make use of tracing, but need not. Indeed, the ability of RS accounts to make sense of the responsibility of cases like Kevin (the grumpy professor) and Bill (the homicidal cuckold) suggest that RS accounts might yet have some decisive advantage over Reasons accounts. That is, RS accounts might have a particularly effective way of accounting for responsibility attributions if it turns out that tracing is as problematic as the above argument suggests.

At this point, matters are too complex to permit any sweeping claims. What we should say partly depends on how serious the tracing worries turn out to be. If they are very serious, then it seems to open some space for the possibility that RS theories have appropriately identified the correct locus of concern for moral responsibility. By focusing on the agent's relationship to the action, as given by the presence or absence of a particular psychological structure (identification, say) it will be less of a concern whether the more general rational capacities that make such psychological phenomena available are frequently operating, engaged, or otherwise immediately present in decision-making. However, if one regards manipulation or implantation scenarios as particularly problematic for RS accounts, or if one were moved by the thought that RS agents can be unacceptably detached from what reasons there are in the world, then one might instead begin to take seriously the prospect of a distinctive form of moral responsibility skepticism. On such a view, one might not think that responsibility is altogether impossible—only much less frequently present than our ordinary practices would suggest.[22]

Here, though, I think the Reasons theorist should resist capitulating too quickly to either the RS view or the attenuated skepticism just mentioned. One reason to think that tracing's troubles are not particularly dire in the present context is that, plausibly, even habitual, personal policy-dictated actions can be sensitive to reasons.[23] That I habitually empty my pockets on a bookshelf when I get home from work does not preclude the following: were there something I perceived as more important, I would respond to those considerations. I do not wish to deny that our habits, traits, or policies can make us less able to detect relevant considerations, moral or otherwise. At the same time, we do well to acknowledge that those same mechanisms can enhance our responsiveness to considerations. If I had no habit of asking my children how their day went, I would presumably fail to be aware of some considerations that should weigh in my deliberations at least some of the time. So, while habit, traits, and policies might sometimes diminish our appreciation for some reasons, they can also work to make us more aware of these things than we might otherwise be. All of which is to say that we should not so readily accept that our reasoning capacity is paralyzed, even when it is silent in action production.[24]

Where does all of this leave us? Answer: with more philosophy to do. In the literature on moral responsibility, RS and Reasons views are frequently treated as offering substantially different approaches to moral responsibility. My aim here has been to show that their relationship is considerably more complicated, but in ways that do not generally favor RS views. Moreover, the recent appeal to experimental data by proponents of RS views is insufficient to address these worries.

None of this is to deny that Reasons accounts face difficulties of their own.

Reasons accounts must explain how we can have adequate sensitivity to reasons in cases where reasons seem to play no active role in the production of action, but where we nevertheless find ourselves inclined to assign responsibility. I have gestured at some initial considerations why one might think that a Reasons account could meet this challenge, but more needs to be said. Whatever is the case about those speculations, however, the difficulties facing Reasons views are comparatively less daunting than those faced by RS views. In particular, it seems that Reasons views face difficulties with tracing more sharply than do RS view *only* to the degree to which RS views make themselves susceptible to manipulation concerns. It is very difficult to see how worries about manipulation cases can be addressed without appealing to tracing or something similar. So, RS theories are left with both the old and the new: familiar and difficult issues with manipulation cases, but also the new challenge of showing that what appeal there is to RS views is not symptomatic of our deeper commitment to the primacy of reasons.[25]

Notes

1. In claiming this, I mean to bracket the larger debate about whether moral responsibility requires indeterminism. That debate is obviously an important one, but tangential for what follows. If indeterminism is a requirement for moral responsibility, we will still need some account of what other features of agency are required for moral responsibility, and it is those other features that are the subject of the present discussion. Also, in what follows I will set aside what is arguably a third alternative in debates about responsibility—"attributionist" accounts of the sort given in T. M. Scanlon, *What We Owe to Each Other* (Cambridge, Massachusetts, and London, England: The Belknap Press of Harvard University Press, 1998); Angela Smith, "Responsibility for Attitudes: Activity and Passivity in Mental Life," *Ethics* 115 (2005): 236-71.

2. Philosophers have given various treatments of what it means to identify with the motives with which one acts, e.g., to be satisfied with the motivating desires, to view those desires as expressing one's true self or true values, and so on. Variants of this picture, broadly construed, has bee suggested in the work of, for example, David Hume, *A Treatise of Human Nature,* trans. L. A. Selby-Bigge, and P. H. Nidditch, 2 ed. (New York: Oxford University Press, 1978); Harry Frankfurt, "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* 68, no. 1 (1971): 5-20; Gary Watson, "Free Agency," *Journal of Philosophy* 72, no. 8 (1975): 205-20; Gerald Dworkin, *The Theory and Practice of Autonomy* (New York: Cambridge, 1988); Michael E. Bratman, "Identification, Decision, and Treating as a Reason," *Philosophical Topics* 24, no. 2 (1996): 1-18.

3. Susan Wolf, *Freedom Within Reason* (New York: Oxford University Press, 1990).

4. I have in mind views of the sort expressed in, for example, Susan Wolf, *Freedom Within Reason*; R. Jay Wallace, *Responsibility and the Moral Sentiments* (Cambridge, Mass.: Harvard University Press, 1994); John Martin Fischer, and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility* (New York: Cambridge University Press, 1998); Nomy Arpaly, *Unprincipled Virtue* (New York: Oxford, 2003); Dana Nelkin, "Responsibility and Rational Abilities: Defending and Asymmetrical View," *Pacific Philosophical Quarterly* 89 (2008): 497-515 and depending on some important interpretive details, it may include such accounts as Robert Kane, *The Significance of Free Will* (Oxford: Oxford, 1996) and Michael McKenna, "The Limits of Evil and the Role of Moral Address," *Journal of Ethics* 2, no. 2 (1998): 123-42.

5. Harry Frankfurt, "Freedom of the Will and the Concept of a Person." (Henceforth: FWCP.)

6. See, for example, some of the varied uses to which Frankfurt's account has been put in James Stacey Taylor, ed. *Personal Autonomy: New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy* (New York: Cambridge University Press, 2005). See also Michael E. Bratman, "Autonomy and Hierarchy," *Social Philosophy and Policy* 20, no. 2 (2003): 156-76.

7. Compare Susan Wolf, "Sanity and the Metaphysics of Responsibility," in *Free Will*, ed. Gary Watson (New York: Oxford, 2003). Wolf uses the notion of 'sanity' in a somewhat idiosyncratic way, but the general thrust of her argument, as I understand it, is to emphasize how those psychological structures that constitute "real selves" require further supplementation by something akin to a reasons condition.

8. This objection was first made in Gary Watson, "Free Agency."

9. Indeed, there is more that Frankfurt went on to say. But in those papers that followed FWCP, the machinery of desiderative hierarchies were re-purposed to account for other agential phenomena (identification, whole-heartedness, and so on) and the matter of responsibility disappeared. In a personal conversation in 1999, Frankfurt said that his views about the requirements for moral responsibility had not changed since FWCP, which further suggests that the work of those later hierarchical accounts, in which 'moral responsibility' virtually never occurs, are not intended as replacements of the earlier account of moral responsibility. So, what follows here are thoughts on how a RS theorist might try to address worries about the account as an account of responsible agency, using the resources of FWCP.

10. There are other ways one might build a RS account. As previously noted, one could appeal to the role of an agent's values, or of the agent's valuings, as part of an account of what constitutes the agent's real self. Indeed, see Gary Watson, "Free Agency." for an attempt to explain how one might answer the challenge he put to Frankfurt without giving up on what I have been calling a RS picture. Alternately, one could appeal to an agent's self-governing policies and their role in securing cross-temporal identity of the agent. See Michael E. Bratman, "Identification, Decision, and Treating as a Reason." It is beyond the scope of this paper to address all possible ways of defending a RS account. Here, I can only flag my suspicion that analogs of several of the already mentioned concerns can be brought to bear against these accounts when they are construed as accounts of responsible agency. But this is not an argument—only an acknowledgement that the most I can hope to show is how reflections on *one* RS theory leads us to a better appreciation of some complexities obscured by the familiar RS/Reasons distinction.

11. For discussions that take RS views, under one or another name, to be an important alternative to what I have been calling Reasons approaches, see, for example: Susan Wolf, *Freedom Within Reason*; John Martin Fischer, and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*; Elinor Mason, "Recent Work on Moral Responsibility," *Philosophical Books* 46, no. 4 (2005): 343-53..

12. John Doris, and Stephen Stich, "As a Matter of Fact: Empirical Perspectives on Ethics," in *The Oxford Handbook of Contemporary Philosophy*, ed. Frank Jackson, and Michael Smith (Oxford: Oxford, 2005); Robert L. Woolfolk et al., "Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility," *Cognition* 100 (2006): 283-401.

13. This scenario was tested precisely because of concerns that in less coercive versions of the case, there remained alternative possibilities that might fuel an incompatibilist reading of the evidence.

14. Doris and Stich explicitly use these results to argue against the intuitiveness of incompatibilism, both of the alternative possibilities variety but also of the variety that does not require alternative possibilities—what Michael McKenna has helpfully dubbed "source incompatibilism." Even if one thought that the evidence cuts against alternative possibilities accounts, I do not see how these data get traction against source accounts. According to source incompatibilists, the removal of alternative possibilities does not, by itself, mean that the agent wasn't the ultimate source of the action. It is difficult to see how one might be an ultimate source without alternative possibilities, but this is precisely the lesson that some source incompatibilists have tried to draw from Frankfurt-style counterexamples to the Principle of Alternative Possibilities (see, Derk Pereboom, "Defending Hard Incompatibilism," *Midwest Studies in Philosophy* 29, no. 1 (2005): 228-47.). Conceivably, a source incompatibilist might argue that Bill was ultimately responsible (assuming he wasn't subject to causal determinism), and that his ultimate responsibility was not

gotten rid of simply because he lacked alternative possibilities. Of course, a source incompatibilist would need some account of what Bill's sourcehood consists in, but I do not see any obvious reason why the case of Bill prevents source incompatibilists from offering an account compatible with the case as it has been described. (Compare the case they rely on with one where the hijackers give Bill a pill that deterministically makes him identify with whatever action they give him. If Bill didn't previously identify with the action, this sort of coercion strikes me as undermining source-hood. I wager it would also undermine the rate at which respondents attribute moral responsibility.) Moreover, there is no reason a source incompatibilist could not help him or herself to a tracing approach (see next section), and thus dodge the consequences of the Woolfolk, et al., evidence in this way.

[15] Kane has proposed a picture along these lines. See Robert Kane, *The Significance of Free Will.*

[16] This aspect of agency plays an important role in much of Michael Bratman's work. See, for example, many of the essays in Michael E. Bratman, *Structures of Agency: Essays* (New York: Oxford University Press, USA, 2007).

[17] Versions of it can be found in various places, both explicitly and implicitly. See, for example John Martin Fischer, and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*; Robert Kane, *The Significance of Free Will*; Laura Waddell Ekstrom, *Free Will: A Philosophical Study* (Boulder, Colorado: Westview Press, 2000); Peter Van Inwagen, "When is the Will Free?," in *Philosophical Perspectives, 3, Philosophy of Mind and Action Theory, 1989*, ed. James E. Tomberlin (Atascadero, CA: Ridgeview, 1989). Versions of a tracing principle also figure prominently in some skeptical arguments, including those in Gideon Rosen, "Skepticism About Moral Responsibility," *Philosophical Perspectives* 18 (2004): 295-313; Galen Strawson, "The Impossibility of Moral Responsibility," *Philosophical Studies* 75 (1994): 5-24.

[18] In contrast, without appealing to tracing, an RS view could say: Kevin is responsible for insulting his students precisely because his doing it is something that expresses his RS (i.e., that he identifies with, that he endorses, that expresses his regard for students, etc., etc.). Note: an RS view may appeal to tracing, but it is not obvious that it must. Or, at any rate, if it must, it need not do so very often. I flag this issue here because it is returns in a later section.

[19] See P. F. Strawson, "Freedom and Resentment," *Proceedings of the British Academy* XLVIII (1962): 1-25.

[20] A further reason for caution about the evidence invoked by Doris and Stich hinges on a complexity of responsibility attributions. In a different set of experiments, Nichols and Knobe discovered that responsibility attributions are sensitive to the way a case is framed. See Shaun Nichols, and Joshua Knobe, "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions," *Nous* 41, no. 4 (2007): 663-85; Shaun Nichols, "Folk Intuitions on Free Will," *Journal of Cognition and Culture* 6, no. 1 & 2 (2006): 57-86. In concrete, high affect contexts, will ascribe responsibility even if they are told it happens in a deterministic scenario. However, when a case is discussed abstractly, in low affect terms, responsibility attributions become much more sensitive to disruption because of determinism (i.e., in high affect contexts, responsibility attributions are resilient in a way they do not tend to be in lower affect contexts.) And, in the Woolfolk et al. experiments, the cases are described in concrete, high affect ways. So, there seems to be a further variable here that needs to be disentagled from their results.

21. Manuel Vargas, "The Trouble With Tracing," *Midwest Studies in Philosophy* 29, no. 1 (2005): 269-91. For a recent reply to these worries, however, see John Martin Fischer, and Neal Tognazinni, "The Truth About Tracing," *Nous* 43, no. 3 (2009): 531-56.

22. To be sure, there are other, independent reasons for worries about the viability of Reasons views (and, correspondingly, RS views if they indeed collapse into Reasons views) in the face of growing experimental data about the production of human action. See, for example, the worries raised about Reasons views in Dana Nelkin, "Freedom, Responsibility, and the Challenge of Situationism," *Midwest Studies in Philosophy* 29, no. 1 (2005): 181-206; Maureen Sie, and Arno Wouters, "The Real Challenge to Free Will and Moral Responsibility," *Trends in Cognitive Sciences* 12, no. 1 (2008): 3-4; Joshua Knobe, and Brian Leiter, "The Case for Nietzschean Moral Psychology," in *Nietzsche and Morality*, ed. Brian Leiter, and N Sinhababu (New York: Oxford, 2007). Elsewhere, I have attempted to address at least some of these worries.

23. Again, I am bracketing compatibilist and incompatibilist disputes about whether one can have unexercised capacities if determinism is true, at least for the purpose of assessing those agential powers central to moral responsiblity. If compatibilism is true, then we will presumably have some way of making sense of what I am saying here, by appealing to something like a counterfactual or dispositional analysis of the capacity. If incompatibilism is true, then we can suppose that what I am claiming is that in cases of habit I retain the relevant libertarian power of rational action-initiation. The latter position would, however, require saying more about skeptical pressures.

24. Tracing might prove to be a more systematic problem for this sort of account if one thought that responsible agency was historical in some deep and systematic sense. Fischer and Ravizza's account of reasons-responsiveness has this feature. They argue that irrespective of what one's reasons-responsive capacities might be, there will always be some historical condition that must be satisfied for one to be a responsible agent. On accounts such as these the historical ownership condition will introduce an element that, at least in principle, seems susceptible to the difficulties that may arise for tracing. But whether and how there is some requirement of history on responsible agency is a complicated matter. For an overview of the relevant literature, see Alfred Mele, "Moral Responsibility and History Revisited," *Ethical Theory and Moral Practice* (forthcoming).

25. Thanks to Daniel Speak for helpful comments on this paper. Thanks also to the Radcliffe Institute for Advanced Study at Harvard, where I worked on this paper.

References

Arpaly, Nomy. *Unprincipled Virtue*. New York: Oxford, 2003.

Bratman, Michael E. "Identification, Decision, and Treating as a Reason." *Philosophical Topics* 24(2), no. 2 (1996): 1-18.

———. "Autonomy and Hierarchy." *Social Philosophy and Policy* 20(2), no. 2 (2003): 156-76.

———. *Structures of Agency: Essays*. New York: Oxford University Press, USA, 2007.

Doris, John, and Stephen Stich. "As a Matter of Fact: Empirical Perspectives on Ethics." In *The Oxford Handbook of Contemporary Philosophy*, edited by Frank Jackson, and Michael Smith, Oxford: Oxford, 2005.

Dworkin, Gerald. *The Theory and Practice of Autonomy*. New York: Cambridge, 1988.

Ekstrom, Laura Waddell. *Free Will: A Philosophical Study*. Boulder, Colorado: Westview Press, 2000.

Fischer, John Martin, and Mark Ravizza. *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge University Press, 1998.

Fischer, John Martin, and Tognazinni, Neal. "The Truth About Tracing." *Nous* 43(3), no. 3 (2009): 531-56.

Frankfurt, Harry. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68(1), no. 1 (1971): 5-20.

Hume, David. *A Treatise of Human Nature*. Translated by L. A. Selby-Bigge, and P. H. Nidditch. 2 ed. New York: Oxford University Press, 1978.

Kane, Robert. *The Significance of Free Will*. Oxford: Oxford, 1996.

Knobe, Joshua, and Brian Leiter. "The Case for Nietzschean Moral Psychology." In *Nietzsche and Morality*, edited by Brian Leiter, and N Sinhababu, 83-109. New York: Oxford, 2007.

Mason, Elinor. "Recent Work on Moral Responsibility." *Philosophical Books* 46 (4), no. 4 (2005): 343-53.

McKenna, Michael. "The Limits of Evil and the Role of Moral Address." *Journal of Ethics* 2(2), no. 2 (1998): 123-42.

Mele, Alfred. "Moral Responsibility and History Revisited." *Ethical Theory and Moral Practice* (forthcoming):

Nelkin, Dana. "Freedom, Responsibility, and the Challenge of Situationism." *Midwest Studies in Philosophy* 29(1), no. 1 (2005): 181-206.

———. "Responsibility and Rational Abilities: Defending and Asymmetrical View." *Pacific Philosophical Quarterly* 89 (2008): 497-515.

Nichols, Shaun. "Folk Intuitions on Free Will." *Journal of Cognition and Culture* 6(1 & 2), no. 1 & 2 (2006): 57-86.

Nichols, Shaun, and Knobe, Joshua. "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions." *Nous* 41 (4), no. 4 (2007): 663-85.

Pereboom, Derk. "Defending Hard Incompatibilism." *Midwest Studies in Philosophy* 29(1), no. 1 (2005): 228-47.

Rosen, Gideon. "Skepticism About Moral Responsibility." *Philosophical Perspectives* 18 (2004): 295-313.

Scanlon, T. M. *What We Owe to Each Other*. Cambridge, Massachusetts, and London, England: The Belknap Press of Harvard University Press, 1998.

Sie, Maureen, and Wouters, Arno. "The Real Challenge to Free Will and Moral Responsibility." *Trends in Cognitive Sciences* 12(1), no. 1 (2008): 3-4.

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

Strawson, Galen. "The Impossibility of Moral Responsibility." *Philosophical Studies* 75 (1994): 5-24.

Strawson, P. F. "Freedom and Resentment." *Proceedings of the British Academy* XLVIII (1962): 1-25.

Taylor, James Stacey, (ed.) *Personal Autonomy: New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy*. New York: Cambridge University Press, 2005.

Van Inwagen, Peter. "When is the Will Free?" In *Philosophical Perspectives, 3, Philosophy of Mind and Action Theory, 1989*, edited by James E. Tomberlin, Atascadero, CA: Ridgeview, 1989.

Vargas, Manuel. "The Trouble With Tracing." *Midwest Studies in Philosophy* 29(1), no. 1 (2005): 269-91.

Wallace, R. Jay. R*esponsibility and the Moral Sentiments*. Cambridge, Mass.: Harvard University Press, 1994.

Watson, Gary. "Free Agency." *Journal of Philosophy* 72(8), no. 8 (1975): 205-20.

Wolf, Susan. *Freedom Within Reason*. New York: Oxford University Press, 1990.

———. "Sanity and the Metaphysics of Responsibility." In *Free Will*, edited by Gary Watson, 372-87. New York: Oxford, 2003.

Woolfolk, Robert L., Doris, John, and Darley, John. "Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility." *Cognition* 100 (2006): 283-401.