Philos Stud
DOI 10.1007/s11098-017-0878-5

Responsibility and the shallow self

Samuel Reis-Dennis[1]

Abstract Contemporary philosophers of moral responsibility are in widespread agreement that we can only be blamed for actions that express, reflect, or disclose something about us or the quality of our wills. In this paper I reject that thesis and argue that self disclosure is not a necessary condition on moral responsibility and blameworthiness: reactive responses ranging from aretaic appraisals all the way to outbursts of anger and resentment can be morally justified even when the blamed agent's action expresses or discloses nothing significant about his or her "deep self," judgments and cares, or the quality of his or her will. I argue that the self-disclosure requirement on responsibility overestimates the extent to which our blaming practices and responsibility judgments are responsive to agents as opposed to actions, and that this mistake has the potential to distort both our reactive responses and our understanding of blamed agents' characters.

Keywords Moral responsibility · Blame · Deep self · Self disclosure · Resentment

1 Introduction

Imagine attending the opening of an art exhibition. Passing by a beautiful sculpture, you decide to smash it on the floor, just for the fun of it. The artist, on hand for the event, is understandably furious and expresses anger, shock, and resentment. These emotions are fitting, but they would have been misplaced had you been pushed into the sculpture, had a mad scientist taken control of you and

✉ Samuel Reis-Dennis
srd2389@gmail.com

[1] Philosophy Department, University of North Carolina, Chapel Hill, Caldwell Hall, 240 East Cameron, Chapel Hill, NC 27599, USA

Published online: 09 February 2017

forced you into the act, or had you knocked it over during a violent, involuntary muscle spasm.

What do pushes, mad scientist interventions, and spasms have in common that standard blameworthy actions lack? One reasonable answer is that in all three of the "blameless" cases, the action is not the agent's own in some morally significant way. The scientist manipulation and the physical push, for example, involve alien causes in the form of other agents; the spasm involves the agent's own brain but tells us nothing morally interesting about him.

Now suppose you are defending a friend to an acquaintance. The acquaintance complains that your friend has been acting erratically, and in his defense, you explain that your friend "hasn't been himself" since his divorce. In offering this bit of information, you intend to make an excuse for your friend: you hope that your acquaintance will suspend or temper his negative reactive response to your friend's behavior. Consideration of these two examples illuminates what I take to be the central thought behind self-disclosure views of moral responsibility and blameworthiness[1]: In order to be responsible for an action, the self-disclosure theorist holds, the action must come from the person, his or her self, in some meaningful way. To articulate this sense of action ownership, the deep-self or self-disclosure theorist appeals to the intuition generated by the second example: what separates actions that come from our deep selves, on the one hand, and actions that our bodies cause but do not really come from *us*, on the other hand, is that the former reveal something important about us, reflect our characters, or, at least, the quality of our wills.

In this paper, I canvas various ways of understanding the deep self. In doing so, I try both to isolate some coherent notion of the deep self and to discuss its relationship to the robust sort of moral responsibility associated with reactive responses like blame and praise. In the second section, I conclude that self-disclosure views get something right: we *do* care about how, and to what extent, actions reflect personalities; we are rightly interested in what we can learn about people through their deeds, reactions, attitudes, and beliefs. Nevertheless, I argue that the connection between the deep self and moral responsibility is tenuous and resists easy explanation. In the third and fourth sections I sketch and respond to the most plausible self-disclosure views on offer in the philosophical literature. I conclude that in many cases responsibility and blameworthiness do *not* depend on the agent's actions expressing anything especially significant or "deep"; "shallow," non-disclosive actions that reveal nothing about the quality of an agent's will can be blameworthy, even in a strong reactive sense. I argue that self-disclosure views overestimate the extent to which our blaming practices and responsibility judgments are responsive to *agents* as opposed to *actions*, and that this mistake has the potential to distort not only our understanding of justified blaming practices, but also our conceptions of the "selves" of blamed agents.

[1] This terminology comes from the work of Susan Wolf. See: Wolf (1987, pp. 46–62).

2 The self in everyday life

Philosophers are not alone in their interest in the self.[2] As a society, we are intensely occupied with the ideal of true-self expression. But what is the connection between the deep self and moral responsibility? One can only begin to understand the philosophical import of the deep self after one gets a sense of our everyday interest in it. What is the real self? Why do we care about it?

We invoke the deep self in a variety of situations. The following is a non-exhaustive list, an attempt to put the concept in context. To begin, we might employ deep-self talk when someone's behavior is substantially at odds with what we have come to expect of her, when she "isn't herself." One might think, for instance, that a family member in the late stages of Alzheimer's "isn't there." Some users of drugs prescribed for depression, bipolar disorder, and other psychological problems report feeling that the drugs make them feel "like themselves," while others express the opposite sentiment, professing that the drugs make them feel like different people entirely. We usually take these reports to be definitive.

Not so with a substantial number of cases of recreational drug use in which we also render judgments of loss of self. Whereas it seems clear that dementia, for example, obscures or erases the afflicted person's true self, our judgments about the deep self and recreational drug use are substantially cloudier. "In vino veritas" is, after all, a well-known saying: we often think that drinking, and some kinds of drug use, can actually reveal, rather than obscure, a person's true character. But this is not always true. When a drinker or drug user does something unfortunate, we sometimes make excuses for him, especially if he is inexperienced with the substance: "he wasn't thinking straight…, that wasn't really him," "that's not the guy I know," "that was the liquor talking."

Instances of stress, sleep deprivation, and fear can function similarly. On the one hand, there is the thought that trying conditions can rob us of our agency. As the Snickers' tagline tells us: "You're not you when you're hungry." On the other hand, we sometimes say that stress brings out our deepest or most significant traits. Legendary UCLA basketball coach John Wooden held that "Sports don't build character. They reveal it," expressing a sentiment echoed in the classic war movie trope in which the jaded sergeant dramatically tells the recruits, "you don't really

[2] I use "self," "real self," and "deep self" interchangeably in this paper. I also use the terms "deep-self views" and "real-self views" synonymously with "self-disclosure views." Some theorists who hold, say, quality of will or judgement-based theories of moral responsibility may object to the characterization of their positions as a "deep-self" views, but terminology is not my main concern, as long as my point is clear: I mean to object to any model that holds that responsible and blameworthy action requires the expression or reflection of an agent's judgments, cares, or the quality of his or her will. On Angela Smith's view (See: Smith 2005, pp. 236–271), for example, morally responsible action must reflect an "evaluative judgment," but such judgments can be "one-off," "out of character," and need not be revelatory of an agent's deepest commitments. For Smith, we are only responsible for actions that reflect evaluative judgments or commitments and that involve the sort of rational activity one could be called upon to justify. This is "deep" enough for my purposes here: I will argue that morally responsible behavior can fall short even of this standard, that some blameworthy actions reveal *nothing* of significance about their authors. I am grateful to an anonymous reviewer from *Philosophical Studies* for urging me to clarify this point.

know what you're made of until you're out there in some bunker, with only your brothers, far from Mommy, surrounded by the enemy…." Interestingly, this feeling is compatible with the excusing power of stress: we sometimes judge that stress prompted the expression of a deep and important part of the agent's personality and yet conclude that his responsibility for the resulting action, and certainly his reactive blameworthiness for it, is diminished.

Another set of cases in which we employ the idea of the deep self involves the familiar appeal to an agent's "true" commitments and character. A parent might think that her son really *does* care about being successful deep down but is just lazy; one might warn a friend that a job candidate "is on his best behavior now, but when he gets comfortable his real personality will come out."

This sort of deep-self talk, implying that only certain actions express our deep moral cores, is related to another use of deep-self discourse in which one appeals to a person's circumstances or upbringing in order to explain or excuse his or her behavior. Like the stress case, this kind of deep-self talk can lead in opposite directions. On the one hand, the idea that there is a soul or moral center inside of everyone, but that circumstantial factors prevent it (the *real* person) from shining through, is a familiar one. On the other hand, we often hold that a man is "a product of his circumstances," or that "his upbringing made him the man he is today." Despite their opposition on the question of the nature of the real self, both of these reactions to a rough upbringing can play a role in excusing bad conduct.

Next, there is the injunction to "be oneself," and the related process of "finding oneself." Once one has "found himself," by organically accumulating a set of values and projects, we might encourage him to be himself when he is about to meet someone new or enter a stressful environment; job interviews, first dates, and situations involving peer pressure come immediately to mind. Sometimes, we cannot "be ourselves" around people who are hostile and judgmental, and we appreciate the ways in which our close friends and family allow us to be who we truly are.

Let us now turn more directly toward the relationship between responsibility, "out of character" action, and expression of the self. Imagine a waiter who is normally painstakingly polite finally snapping at the end of a long day of work. The diners unlucky enough to be in the restaurant for his eruption might think they have seen his "real self." But depending on what the waiter says and does, he might turn out to be blameworthy regardless of whether or not he actually revealed anything especially deep or important about his character. If he really was just having a bad day, for example, and the incident was a genuine aberration, he could still be a legitimate target of the reactive attitudes. The general point is that many blameworthy actions are not revelatory of anything persistent and stable in the agent's character, and this is reflected in our ordinary understanding of the deep self.

Of course, this survey of our interest in the deep self it is not exhaustive, but I think it is sufficient to situate the concept in its everyday context. What can be gleaned from this overview? Obviously, we can see that the real or deep self is important to us. The sheer diversity of cases above indicates that we are, across a wide array of settings, deeply concerned with real-self expression.

But perhaps the cases are so diverse that they resist systematization. Is there just one notion of the deep self underlying all of this use? Certain examples suggest fragmentation. Recall the case of stress: some say that stressful situations obscure the "real" self, while others claim such situations reveal it. In fact, I am optimistic that both camps are, at least in most cases, out to capture and describe the same thing. At the very least, a vast majority of these examples suggest some notion of a person's core—a stable, enduring, personality.

The relationship between this important everyday conception of the deep self, however, and judgments of responsibility and blameworthiness, is less clear. It seems to me that much of our interest in the deep self is concerned not with questions of agency, or the fittingness of praise and blame for actions, but rather with notions of identity and authenticity. Consider again the command to "be yourself," and imagine that a man on a first date becomes nervous and, as a result, ignores the injunction and puts on an elaborate facade of macho posturing, embarrassing himself and insulting his date. His failure to be himself is morally interesting, but this does not imply that his responsibility for his boorish behavior is somehow mitigated. Even on the assumption that his actions really were out of character and explicable in part by his nervousness, his date might still reasonably feel disappointed, hurt, and angry with him on the basis of his display. At the same time, it isn't as if learning that the man's actions did not reflect his deep self would mean nothing. The revelation might prompt the woman to give the man another chance, or at least to moderate her lasting ill will toward him. On the other hand, it also suggests a failure of authenticity that she might take to be problematic in its own right. Of course we care a great deal about what particular actions reveal about who we are deep down; we form self-conceptions and pursue and mold relationships in large part on the basis of judgments about what our actions and the actions of others mean. But being "true to oneself" is sometimes hard, and we often fail. In such lapses of authenticity, we are still accountable for our actions, even when they genuinely do not express something "deep" about our personalities and characters.

Obviously, the cases of the waiter and the failed date will not refute all varieties of self-disclosure theories. (Theories that emphasize quality of will, for example, can account for them rather easily.) There are various philosophical moves to be made in discussion of examples like these, and I will rehearse and address them in the next two sections. My point here is that philosophers who wish to explain responsibility in terms of deep-self disclosure will *need* to make some moves, and that these moves will put a distance between their theorized conceptions of the deep self and the concept of the deep self we employ in everyday life.

Self-disclosure theorists have failed to fully appreciate this disconnect between the nature of our everyday interest in our deep selves and judgments of moral responsibility for individual actions. In what follows, I will argue that the *responsible* self is actually quite shallow, in some cases almost empty.

3 Frankfurt and Watson

In the previous section, I set out to show that the ordinary sense of the deep self, though important, is not systematically connected to responsibility judgments. Expression of the morally important concept of the deep self, as it is understood in everyday life, is not necessary for responsibility and reactive blameworthiness. In this section I will canvas and evaluate some more highly theorized notions of the deep self that philosophers have tried to link directly to responsibility and blameworthiness.

The logical starting point in a survey of contemporary attempts to isolate the responsible self is the work of Harry Frankfurt[3] and Gary Watson.[4] While Frankfurt and Watson both tie their conceptions of the real, authentic self to moral responsibility, making this connection is not the only purpose of their work. Unlike some more recent deep-self theorists I will come to later, Frankfurt and Watson intend, it seems, to capture something close to the intuitive notion of the deep or authentic self that I articulated in the previous section.

Frankfurt is particularly interested in the concept of a person, arguing that our interest in our wills and our capacity for reflective self-evaluation are the defining features of personhood.[5] According to Frankfurt, this caring about what we do manifests itself in the form of second-order desires, or desires about desires. One kind of second-order desire is a second-order volition. This is a desire for a first-order desire to be effective in action, a desire that we act on some first-order desire. Frankfurt holds that only persons have second-order volitions. The wanton, a man with first-order desires but no second-order volitions, does not care about his will.[6] Genuine persons are not like this. Only persons can be free and responsible, and in free action we exercise our personhood by acting on the desires we want to be effective.[7] To see how this works, consider the case of an unwilling drug addict. This man has a first-order desire to get high and a first-order desire to stay sober. He wants his desire to stay sober to win out, but he is overcome by his desire to get high. This person, Frankfurt explains, does not act in accordance with his second-order volition and is thus unfree.[8]

Watson's view is structurally similar to Frankfurt's. He argues that we are free and responsible when we act in accordance with our values, questioning whether we need to stipulate anything about second-order desires and volitions in order to know that certain voluntary actions are *ours*.[9] What is special about our agency, Watson explains, is our capacity to value things, to judge that they are good or worth doing. When we express these judgments and act on our values, we are free and

[3] Frankfurt (2003, pp. 322–336).

[4] Watson (2003, pp. 337–351).

[5] Frankfurt, p. 329.

[6] Ibid, p. 327.

[7] Ibid, p. 330.

[8] Ibid, p. 331.

[9] Watson, 'Free Agency', p. 350.

responsible. What difference does it make, he wonders, whether we have a second-order desire for some first-order desire to be effective?[10] As long as that first-order desire is a desire for something we value, then we are free and responsible when we act on it.

These views have been discussed in great detail in the responsibility literature, and so I will try to keep my remarks brief. A few points are worth emphasizing. First, as I said at the outset, both theories are of great interest even if their verdicts on responsibility are implausible. Both accounts attempt to explore something close to the ordinary conception of the deep self, which, as we have seen, is central to judgments of identity and authenticity. Still, I am not the first to note some drawbacks of these views when it comes to their ability to explain our intuitive judgments of responsibility and blameworthiness. One classic problem involves consideration of the role of the historical origins of the self-constituting desires or attitudes. For example, if one's second-order desires and values are the results of a neurological disorder, a trauma, or even brainwashing, Frankfurt and Watson's theories would seem to render unintuitive, even harsh, verdicts.

Perhaps even more important is the point that rational disavowal or lack of second-order endorsement is not intuitively excusing in a large cluster of everyday cases of blameworthiness. Consider the case of Blackout Bill, a college student who gets intoxicated almost to the point of alcohol poisoning, stumbles into his friends' apartment, and, to everyone's shock, launches into a horrifying racist and sexist rant in which he expresses apparently deeply buried desires to perform unspeakable acts. It seems to me that Bill's friends could accept his insistence that his words do not truly reflect his stable character. They might have no reason to think that Bill actually endorses, or in fact has ever even indulged, any of the terrible thoughts he expressed in his rant. Frankfurt's view, which relies on the presence of second-order desires to ground responsible and blameworthy action, cannot make sense of Bill's blameworthiness. After all, it isn't clear that Bill had *any* second-order desires in the moments leading up to his rant, nor do I think it matters if he didn't. Watson's form of the view is, to my mind, similarly ill-positioned to respond to the case. It is easy to imagine that Bill's actions were genuinely out of character, that he does not reflectively endorse the views he expressed. In other words, it is easy to imagine that Bill's actions do not reflect his "real self," at least in the senses that Frankfurt and Watson understand it. Though this realization might be comforting to Bill's friends, it would still be fitting for them to feel angry, indignant, and hurt by his words, and for Bill to feel guilty and ashamed.

4 Responsibility and the shallow self

Frankfurt and Watson's theories have given rise to a substantial literature on the relationship between the deep self and responsibility. Descendants of Frankfurt and Watson, despite their general agreement that responsibility is a matter of self-

[10] Ibid, p. 350.

disclosure or self-expression, diverge in their identifications of the deep, responsible self. Broadly speaking, one camp, which includes TM Scanlon[11] and Angela Smith,[12] argues that the true self is expressed in those actions that reflect our judgments or rational activity. Of course even among these theorists there are differences. Here, I focus on Smith's account in particular, but I intend for my remarks to apply more widely.

On Smith's "rational relations" theory, moral responsibility is tied to moral address. The hallmark of responsible action is that the agent can be called to account for what he has done. The intelligibility of the demand for explanation and justification, and the obligation to admit fault if these are not forthcoming, are, for Smith, the central features of responsible action. She writes:

> Most of our desires, emotions, beliefs, and other attitudes seem to meet this condition of judgment-dependence, even though they do not commonly reflect a choice or decision, and are not normally under our voluntary control. These states are ''judgment-dependent'' in the sense that they generally reflect and are sensitive to our (sometimes hasty, mistaken, or incomplete) judgments about what reasons we have, and they are generally responsive to changes in these judgments. We are ''responsible for'' these things, therefore, because they reflect rational assessments for which we are appropriately regarded as answerable.[13]

While Smith's view, with its emphasis on rationality and judgment, is in some ways a close relative of the Watson account I discussed earlier, her conception of the responsible self stretches the notion of "judgment." On Smith's view, both cognitive and conative states are accountable to evaluative judgments. If properly connected to evaluative judgment actions, omissions, attitudes, and the traits that give rise to them could all be fair game for moral criticism, praise, and blame:

> What matters, on this account, is whether an action or attitude is normatively connected to a person's underlying judgments in such a way that she can, in principle, be called upon to defend it with reasons and to acknowledge fault if an adequate defense cannot be provided. Bodily movements and mental states that are not even in principle answerable to a person's judgment are therefore not the sorts of things for which we are responsible, on this account; but we are responsible for most of our desires, emotions, beliefs, and other attitudes, despite the fact that they do not generally arise from conscious choice or decision and are not normally under our immediate voluntary control.[14]

Before I respond directly to Smith's version of the self-disclosure view, I will introduce a closely related rival. On the competing theory, the deep self is associated not with an agent's evaluative judgments but rather with his or her

[11] Scanlon (1998).

[12] See Smith (2005, 2008, pp. 367–392).

[13] Smith, 'Control, Responsibility, and Moral Assessment', p. 370.

[14] Ibid, p. 370.

conative states and dispositions. Chandra Sripada, for example, holds that the responsible self is to be identified with the agent's *cares*, understood broadly to involve "a complex syndrome of motivational, commitmental, evaluative, and affective dispositions."[15] On his view, judgments about someone's deep self, and thus about responsibility, are primarily concerned with responding to the sense the agent gives us of what *matters* to him.[16]

In response to these dueling proposals, David Shoemaker has offered what he calls "the ecumenical view" of the real self. On this theory, the deep self is disjunctive: actions that express either an evaluative judgment or a commitment count as coming from our responsible selves and make us legitimate targets of moral appraisal.[17]

I do not want to go too far into the details of any of these theories here. My criticisms do not depend on their intricacies but instead aim to strike at a more fundamental level. I argue that each of these proposals suffers from a common mistake: they assume that responsibility judgments, expression of reactive attitudes, and moral criticism are mostly about assessing a person rather than his or her action, and this is false. Determinations about the extent to which an action reflects an agent's true personality, evaluative judgments, or cares are important to our blaming practices, but self-disclosure is not a necessary condition of blameworthiness.[18] In deciding whether an agent is blameworthy for an action, we care about self-disclosure only because it affects what Scanlon has called the "meaning" of the action (its significance for affected agents),[19] not because it is a necessary condition on moral responsibility. When an action does express or reveal something significant about its author, this is relevant to the content and tone of the blame we direct toward him, but the absence of self-disclosure is not fully exculpating.

Consider instances of forgetting. It is supposed to be a great advantage of self-disclosure theories that they can accommodate our intuition that we are responsible, for example, for forgetting friends' birthdays or to pick one's child up from school.[20] In fact, self-disclosure theories fall short in their treatment of such cases, generating "blameless" verdicts for culpable agents. For now, consideration of Shoemaker's proposal will be most instructive. I will show that even the most "ecumenical" theory of attributability as self-disclosure has the potential to distort our reactive judgments.

Recall that for even the most inclusive self-disclosure theorists, what matters to an agent is central, and that to be blameworthy for an action the action must be

[15] Sripada (2016, p. 1211).

[16] Ibid, p. 1211.

[17] Shoemaker (2015, pp. 115–140). Shoemaker is concerned specifically with aretaic appraisal, but this will not be important to my criticisms, which will show that some actions that express neither an agent's commitments nor his evaluative judgments can be subject not only to aretaic appraisal but even to strong forms of reactive praise and blame.

[18] My arguments are meant to apply to the entire spectrum of self-disclosure views, including, for example, theories like the one offered by Nomy Arpaly, which takes the expression of deficient quality of will to be necessary for blameworthiness. For a defense of this view, see Arpaly (2006).

[19] Scanlon (2008, p. 52).

[20] Smith, in particular, makes significant use of these examples.

evidence that one's commitments, cares, or judgments (ways of mattering) are somehow deficient. The self-disclosure strategy in the forgetfulness cases, then, involves the claim that the parents and friends in the examples are blameworthy because their evaluative judgments about the importance of the birthday or their level of caring about their children open them up to moral criticism.

Surely this is true of some forgetting cases, but need it be true of all of them? Is every blameworthy instance of forgetfulness traceable to some genuinely "deep" shortcoming? I think not. We can easily imagine cases in which one *just forgets* despite genuinely caring deeply, though not obsessively, about a friend's birthday. Or consider the following case: Johnny has agreed to pick up his friend Camila at the bus stop. In the morning, she calls to let him know that she decided to catch an earlier bus because of concerns about a snowstorm that is expected to hit in the afternoon. Right after he hangs up the phone, Johnny gets word of a relative's medical emergency. He spends the day talking to various family members, is totally preoccupied, and, because the snowstorm never hits, is never reminded of the early pickup. A few hours later, he gets a call from Camila and realizes, to his horror, that he has left her stranded outside in the cold!

Is Johnny blameworthy for his forgetting? A self-disclosure theorist might argue that he is, because his forgetting reveals an objectionable lack of concern for his friend. But does his forgetfulness necessarily show anything of significance about his cares, evaluative judgments, or quality of will? Perhaps, in some version of the case, but I don't think we need to answer "yes" to conclude that it would be fitting for Camila to feel, and even express, some level of anger, upon his arrival, *and even after she learns the full explanation of his tardiness*. After all, Johnny has no excuse; no one made him forget to make the pick-up—the mistake was entirely his fault. As a result, he owes Camila an apology despite the fact that the afternoon's events revealed, if anything, only that he has a very specific disposition to forget *this particular thing* under *these* highly unusual conditions.

At the most general level, self-disclosure theories make the mistake of positing a strict metaphysical condition on responsible and blameworthy action that does not reflect the actual norms of justified blaming practices; thus, when the condition is not met, the theory automatically generates blameless verdicts for blameworthy agents. To more accurately assess Johnny's blameworthiness, we may bypass the metaphysical responsibility conditions checklist and proceed directly to the level of moral response. My sense is that it is fitting for Camila to be upset upon Johnny's eventual arrival because *blame and indignation are often responsive to agents' failures to fulfill their obligations or meet reasonable expectations*. Johnny has let his friend down. He has failed in a duty; to go on without an apology or an expression of remorse would be to devalue Camila, to send a message that he is socially superior, above reproach. And it is not merely that it would be *nice* of Johnny to apologize; it is, rather, that an apology is *owed*: even though his omission doesn't necessarily reveal anything "deep" or interesting about him (his judgments, his cares, the quality of his will, etc.), he has wronged his friend. He made a mistake and he's accountable for it, just as he would have deserved praise, or at least a "thank you," had he executed his duty as promised.

Consideration of guilt and apology in these contexts can help to bring out the point. When we let down people we care about because of a failure or mistake that is entirely ours, feelings of guilt can reveal an investment in the relationship, a fitting discomfort in knowing that our error has harmed another, and a desire to do better in the future. The expression of these feelings can be significant for a relationship after this kind of non-disclosive failure or mistake; in the absence of an apology, the wronged party may feel that he is diminished, that his pain has been forgotten or overlooked, and may be justified in employing angry blame as a means of eliciting a gesture of remorse.

To be clear, this does not mean that the lack of self-disclosure is irrelevant in such cases. Quite the opposite: the fact that Johnny *does* care a great deal that his omission led to his friend's being marooned at the bus stop in the cold is something he would no doubt emphasize when he finally arrived. All I am objecting to is the idea that the non-revelatory nature of the forgetting in this case gets Johnny off the hook completely. It does not. He really *does* care about stranding Camila, and this is precisely what makes him feel (appropriately) remorseful and guilty, two hallmarks of blameworthy behavior. Often, angry feelings in these kinds of cases will be short-lived (especially if an apology is offered), although not always. After all, the consequences (and a host of other factors) also affect the meaning of the action and the kind of reactive response that is warranted: if Camila gets frostbite waiting for Johnny to pick her up, then her justified blaming reaction may be harsher because of the event's greater seriousness and significance.

But, a self-disclosure theorist may object, even if one grants that Johnny's forgetting does not reflect a judgment that Camila isn't important, surely it *does* reflect a judgment about the significance of his relative's medical emergency.[21] And because the forgetting is traceable to *this* judgment, and the resulting preoccupation, the self-disclosure theorist may argue that her theory can account for Johnny's blameworthiness after all.[22]

It is worth noting that this response trades on a subtle but substantial shift on the part of the self-disclosure theorist. So far, we have been working under the assumption that on the self-disclosure model, blameworthiness requires the expression of an *objectionable* or *deficient* evaluative judgment.[23] In the case at issue, however, all of Johnny's judgments (assuming, as we have been, that he cares both about his relative's situation and about Camila's welfare) are appropriate, and even commendable. There is no deficient evaluative stance, attitude, or judgment upon which to ground his blameworthiness.

[21] For simplicity, I will focus on judgments in discussing this objection, but the response applies to care-based views as well.

[22] I am grateful to an anonymous reviewer from *Philosophical Studies* for raising this important objection.

[23] Smith seems to suggest this version of the view. She writes, for example, of someone whose contempt for members of a racial group stems from a judgment about their intellectual inferiority that "the person is open to rational, and in this case moral, criticism for this attitude precisely because of its rational dependence on these objectionable underlying evaluative judgments." ('Responsibility for Attitudes: Activity and Passivity in Mental Life', p. 254).

But even a more relaxed version of the self-disclosure theory, on which blameworthy action may be traceable to *admirable* evaluative judgments, will struggle to explain a range of forgetting cases. Imagine, for example, a slightly different iteration of the Johnny and Camila example: After Camila calls to change the pickup time, Johnny receives an email from a coworker linking to a mildly amusing YouTube clip rather than a phone call about a sick relative. One video leads to another and, eventually, Johnny loses track of the time. From here, everything else is the same: the snow storm never hits, he isn't reminded of the early pickup, he strands Camila.

Is the forgetting in this modified example traceable to an evaluative judgment of Johnny's? My contention is that it need not be for him to be responsible and blameworthy for his lapse. For Smith, "[Evaluative] judgments, taken together, make up the basic evaluative framework through which we view the world. They comprise the things we care about or regard as important or significant."[24] But to claim that anything of even moderate significance is necessarily disclosed in this case seems overly hasty. In the modified example, what disturbs Johnny's routine is a sequence of random, banal distractions, combined with the change to the pickup schedule he had originally planned for. His forgetting need not be traceable to any evaluative judgment at all. Left without a judgment of even fleeting significance to stand on, the self-disclosure theorist, should she wish to take this route, may be forced to conclude that Johnny's blameworthiness for his forgetting in the YouTube version of the case is somehow *more* open to question than it is in the original story, in which he is distracted because of news that he rightly cares deeply about. This would be an odd result.

At this point, one may be tempted to respond that Johnny's forgetting, especially in this newest iteration of the case, must show something about him after all. Could someone who truly cared about his friend have zoned out watching internet videos? In short, yes. I contend that these sorts of non-revelatory lapses are indeed possible (and even relatively common!), and that we may be rightly blamed for them when they occur.

At the outset, I suggested that self-disclosure views can distort our understanding of blamed agents. I am now in a position to explain myself more fully. As we saw in section II, we are deeply invested in the everyday notion of the self. The idea is central to our forming self-conceptions, to understanding our relationships with others, and to shaping the narrative structures of our lives. Those who hold self-disclosure views of responsibility risk warping these processes. I have suggested that the forgetful person might be blameworthy without revealing anything significant about his evaluative judgments or his cares. I have stressed the oddness of giving up on blame in cases like these, but saving the reactive attitudes by insisting that these lapses *are* self disclosive after all may be even worse. We care a great deal about what actions tell us and don't tell us about actors, and to insist that all blameworthy mistakes, reactions, lapses, and failures of memory reveal something deep about us, our judgments, or the quality of our wills could lead to

[24] Smith, 'Responsibility for Attitudes: Activity and Passivity in Mental Life', pp. 251–252.

potentially significant distortion; the more invested in the importance of the self[25] we are, the warier of the inference from blameworthiness to "deep" deficiency we ought to be. To claim that Johnny's blameworthy lapse then, need be revelatory of his judgments, cares, or quality of will, is to invite a mistaken judgment about his character (or, at least, his judgments/cares). He is responsible and blameworthy for his forgetting because he stranded his friend and has no good excuse; what the incident actually reveals of his judgments, cares, and character is an important further question that deserves to be investigated and answered on its own terms.[26]

These cases are closely related to another group of counterexamples to self-disclosure views of this kind: mistakes. It may be helpful to think first about a recent non-moral example: When Minnesota Vikings kicker Blair Walsh missed a short field goal attempt that would have sent the Vikings to the second round of the 2015 NFL playoffs, he said: "It's my fault…. I'm the only one who didn't do my job there. So that's on me…. I worked real hard to get myself to a place where I was very consistent for this team all year, and in that moment, the moment they needed me the most this year I wasn't and that stings…. I'll take the blame because I deserve every second of it."[27] Is Walsh guilty of any failure of caring, evaluative judgment, or quality of will here? It seems unlikely. After all, he seems to care deeply about the team and to have had an excellent work ethic that reflects his exemplary level of investment. One could also claim, I suppose, that his miss was a failure of football-related evaluative judgment. While this may be true of some version of the case, we can easily imagine that no evaluative judgment was involved at all, that Walsh simply acted. *He just missed*. Still, despite his purity of evaluation, commitment, and will, Walsh's teammates and coaches might reasonably be upset with him: he should have made the easy kick, and his misfire cost his team the game. As he put it in his postgame remarks, he didn't do his job, and, as a result, he must "take the blame."

We can construct similar, graver, cases on the same model. A doctor tries to do her very best but makes a simple mistake during surgery. An accountant, even after a thorough double check, makes an oversight that costs a coworker her job.[28] These

[25] Or our judgments, cares, etc.

[26] This is not to say, however, that the two questions are *entirely* independent. As I have said, the sort of blaming reaction that will be fitting in response to Johnny's forgetting will vary depending on how revelatory of bad character the incident is. What I am objecting to is the claim that disclosure of anything significant is required to cross the blameworthiness threshold. I am grateful to Vida Yao for encouraging me to discuss the way in which the internalization of a self-disclosure view can lead to mistaken character judgments.

[27] See: 'Walsh Shoulders Blame for Devastating Loss' (2015).

[28] In response to this case, a Smith-style self-disclosure theorist may note that the accountant in question *does* make a faulty judgment of a certain kind: she should not have judged, say, that the books were balanced when in fact they were not, or that the paperwork was suitable for submission. While this is one way of describing the situation, it is not in the spirit of Smith's use of the term "evaluative judgment." Smith is focused on those judgments that one can, in principle, be called upon to justify and that involve one's "'taking' or 'construing' things to be a certain way or to have a certain significance" (Smith 2005, p. 260). To interpret clerical errors as involving such judgments seems a stretch; it would be odd to ask someone to justify his overlooking a number on a spreadsheet, or his mistakenly writing that the sum of two and two is five. It *does* make sense to ask one to justify his decision not to double-check his work, but

agents, given their roles, are obligated to do their jobs correctly. The fact that their errors do not reflect misguided values, lack of caring, or substandard quality of will means that their actions aren't as bad as they could be; nevertheless, it is fitting when people they harm feel and express some degree of anger toward them.

I think these cases show that ensuring even that weak conditions of responsibility are satisfied before determining how we ought to respond to an agent's behavior can lead to confusion and generate dubious results.[29]

The alternative requires, first, that we skip the "responsibility conditions checklist" altogether when confronted with an action, and move directly to a discussion of its meaning, cognizant of its context and consequences, and of the proper response, informed by an understanding of our reactive practices. On this method, the way to formulate a thesis of self-disclosure would not be to say, simply, "moral responsibility depends on self-disclosure," but, rather, "in order to let someone down or violate an obligation, the agent's action must be self-expressive." And this thesis, as the cases I have been discussing demonstrate, is false: we do not need to reveal anything of significance about our judgments, cares, or the quality of our wills to disappoint others and fail to meet reasonable expectations.

For the self-disclosure theorist, as we have seen, these cases pose a dilemma: either the forgetting, erring, or failing person betrays a deficiency of caring, judgment, or quality of will, or else she is not blameworthy for her behavior and its results. I have rejected this dilemma, but I should say a bit more about the theoretical underpinnings that give it its intuitive appeal.

Philosophers of moral responsibility have found themselves in this position, I think, because of an antecedent attachment to a theoretical picture on which blame and responsibility depend entirely on metaphysical or psychological features "internal" to the blamed agent, and which suggests a theory of blameworthiness on which an impartial observer doles out demerits in perfect proportion to the "stains" on the agent's soul.[30] While many contemporary self-disclosure theorists would no doubt denounce such a conception of blame and blameworthiness, one can see the remnants of this picture in almost every prominent theory on offer in the contemporary literature. The relic of this understanding of blameworthiness is the

Footnote 28 continued
I have been pointing out that mistakes can occur even when one *does* double-check, and arguing that we are often still responsible for them when they do. Thank you to an anonymous reviewer from *Philosophical Studies* for pressing me to clarify this point.

[29] But without conditions of responsibility, how are we to distinguish between types of failures? There is a difference, after all, between Johnny stranding Camila because he forgot to pick her up and stranding her because someone slashed his tires; surely Walsh would get a pass for his missed kick if, at the last second, he had been swarmed by bees and stung in the leg. A procedure for distinguishing blameworthy from non-blameworthy cases of non-disclosive failure would require nothing short of a new theory of moral responsibility. And while I think that one must take account of the arguments and examples I have discussed in this paper in developing a theory of this kind, I cannot undertake that formidable task here. I will, however, in what follows, suggest a way in which we might fruitfully adopt a more social or "ecological" model in our thinking about moral responsibility and blame. While this sketch will fall short of amounting to a fully satisfying theory, I hope it represents a step in the right direction. I am grateful to an anonymous reviewer from *Philosophical Studies* for raising this question.

[30] For a view of this kind, see: Zimmerman (1988).

self-disclosure condition itself, the presence of which ensures that blameworthiness depends upon, and that blame responds to, some failure of the blamed agent's soul, self, character, judgment, or quality of will.

Of course, in many cases it seems obvious that blame is responsive to the bad features of a wrongdoer's character that are manifested in his or her action. But even in cases in which the expression of ill will is essential to understanding blame, the blaming reaction is shaped and textured by other factors that affect the action's meaning: consequences, relationships, and context.

The implausibility of theories that demand that the harshness of our reactions to bad behavior should perfectly track a blamed agent's level of controlled wrongdoing, or poor quality of will expressed in his or her action, has been apparent to laypersons and philosophers alike. It is immediately obvious, for example, that even the most innocuous uses of "tracing" procedures by which one "traces" an agent's blameworthiness for an action or outcome back to some earlier failing seem to license the expression of harsher reactive attitudes toward unlucky agents than toward their lucky counterparts. In one sense, we might say that two drunk drivers, both of whom demonstrated the same lack of good will and failure of caring and judgment when they got into their cars after a night of drinking, for example, are "equally blameworthy," despite the fact that one driver was unlucky and injured a pedestrian and the other returned home without incident. But when it comes to the responses sanctioned by our moral practices, the character and tone of the reactive blame that is appropriately directed toward the drivers is subject to huge variation depending on what happened on their trips home. As Thomas Nagel puts the point in "Moral Luck":

> The same degree of culpability or estimability in intention, motive, or concern is compatible with a wide range of judgments, positive or negative, depending on what happened beyond the point of decision. The *mens rea* which could have existed in the absence of any consequences does not exhaust the grounds of moral judgment. Actual results influence culpability or esteem in a large class of unquestionably ethical cases ranging from negligence through political choice.[31]

If one is willing to concede that reactive response ought to vary at all in this way, I think that one ought also be willing to entertain the thought that requiring even a minor deficiency of will or character in order to get justified blaming reactions off the ground may also be a mistake. In other words, one should be ready to take seriously the contention that the sorts of faults or mistakes that the reactive attitudes address (and ought to address) are not confined to deficient psychological states, or even best described in "internal" terms. One lesson of moral luck is, of course, that our justified responses are shaped in part by the outcomes of blamed agents' behavior. But this insight suggests another lesson that may be even deeper and more important: namely that we are not always interested in reacting to psychological states or motions of the will at all; the reactive attitudes respond to actions at a different level of description.

[31] T. Nagel 'Moral Luck' in Nagel (1979, p. 30).

We react to actions as events in an ecosystem, not as internal phenomena. Interpersonal expressions of resentment and blame are social responses; when we blame, we describe the blameworthy behavior in "thick," inescapably social, terms. One is blamed for hurting another's feelings, for stranding a friend at the bus stop, for botching a surgery, for getting a colleague fired. These are the socially disruptive "actions" that reactive blame is equipped to address.

This is not to say that blameworthiness does not require some *wrongdoing* or *lapse* on the part of the blamed agent. It does.[32] But once we describe wrongdoing in the "thick" social terms I have suggested, we can begin to take a view of blameworthy behavior that is broader than the narrow, "internalized," conception of blameworthy failure imposed upon us by self-disclosure views. This understanding of the ways in which justified reactive practices work can help to explain, not only the persistent judgment that contingent factors like consequences should affect blame, but also why actions (as opposed to mental states or the internal workings of the will) are the loci of blame in the first place.

We are now in a better position to evaluate Smith's claim that "Moral criticism, by its very nature, seems to address a demand to its target. It calls upon the agent to explain or justify her rational activity in some area, and to acknowledge fault if such a justification cannot be provided."[33] Consideration of the foregoing cases suggests that Smith's thesis is true of some forms of moral criticism but does not tell the whole story. Moral criticism, and our blaming practices more generally, are not so straightforward. When we engage in these activities, we respond not only to personality defects, failures of caring or judgment, and lapses in quality of will; we engage in a process that allows us to express a wide range of feelings that are responsive not just to individual elements of the target's psyche or their expression in action, but to the whole state of affairs those actions gave rise to. These responses allow us to stand up for others and ourselves, restore social order, articulate standards of conduct, and, if things go well, to start an interaction that could prompt the target to feel the self-inflicted pain of guilt and offer an apology. The blaming process does not end with, nor does it aim at, a mere acknowledgement of fault, at least in the narrow sense of "fault" that the self-disclosure theorists I have discussed seem to have in mind.

Before concluding, I wish to address one more variant of the deep-self approach suggested by Timothy Schroeder and Nomy Arpaly.[34] Schroeder and Arpaly offer what they call a "whole-self" view in response to hierarchical deep-self views that privilege one aspect of the self (rational judgment, for example) over another (say, desire). Many of the arguments I have offered here apply to the whole-self view, as

[32] For this reason, the point is distinct from Bernard Williams's discussion of the "lorry driver who, *through no fault of his*, runs over a child," (my emphasis), and feels agent-regret, or Adam Smith's consideration of "a man of humanity, who accidentally, and without the smallest degree of blamable negligence, has been the cause of the death of another man, [and] feels himself piacular, though not guilty." In the cases I am interested in, there *is* fault, and so guilt and blame, rather than agent-regret, are fitting. See: Williams (1981, p. 28), and Smith (1976, p. 107).

[33] Smith, 'Control, Responsibility, and Moral Assessment', p. 381.

[34] Arpaly and Schroeder (1999, pp. 161–188).

it is in some ways quite similar to Shoemaker's "ecumenical" proposal. But in their discussion of the whole self, Schroeder and Arpaly make some interesting remarks on the integration of the self that merit additional discussion.

First, this claim: "Other things being equal, an agent is more praiseworthy for a good action, or more blameworthy for a bad action, the more the morally relevant psychological factors underlying it are integrated within her overall personality."[35] In many cases, this seems plausible. Consider once more the case of ranting Blackout Bill: Bill does not reflectively endorse racist and sexist thoughts, and I argued that his high level of blameworthiness is, in large part, due to bad luck. But what if Bill frequently indulged the fantasies he expressed in his diatribe and secretly resented women and racial minorities? It seems to me that this second version of Bill might merit harsher blaming reactions than his unlucky counterpart, and it is plausible to explain the difference in fitting reactive response by appealing to the way the rant fits with, and gives voice to, the racist and sexist elements of Bill's stable character.

In other cases, though, the integration principle seems to fail. Imagine two men who conspire to defraud patrons of local restaurants by stealing their credit card numbers. Neither man needs the money, and both have committed this sort of crime countless times before. In fact, the only difference between the two men is to be found in the thoughts that run through their heads as they commit their crimes. The first man experiences no internal conflict and always executes the job smoothly. The second man is always struck by the hardship he is about to inflict upon the local restaurant owners and their employees every time he starts to steal the credit card numbers. Then, the thought passes and he executes the job smoothly. Who is more blameworthy for the fraud? The answer is not at all obvious. To my ear, the most plausible reading of the case is that the first man is the worse man, but that the second man's actions may be deserving of a harsher reactive response. After all, the fragmentation of his psyche forced him to confront the suffering he was about to cause…and he did it anyway! At the very least, the matter should be up for discussion. It is far from obvious that the more divided man deserves less blame than his wholehearted friend. In fact, there is no reason to assume even that the two merit the same kind of reaction.

Certain cases of praiseworthy actions cause a similar problem. Consider the case of Huckleberry Finn's decision not to turn Jim in that is central to Schroeder and Arpaly's argument against hierarchical views of the self. They write: "Of course, Huckleberry would be more praiseworthy still if he had no racist beliefs at all—if he were more integratedly good—but he may merit high praise nonetheless."[36] Once again, this doesn't strike me as obvious: Perhaps integrated Huck is a better person overall, but the internal conflict of Twain's original makes his *action* especially praiseworthy. In order to do the right thing, he had to choose friendship and compassion over an incorrect, but settled, moral judgment…, and this is impressive! A more integrated Huck's parallel action would be worthy of praise as well, but perhaps not to the same degree, and perhaps not even of the same kind.

[35] Ibid, p. 172.

[36] Ibid, p. 178.

Second, and more directly related to a central theme of this paper, is the Schroeder/Arpaly tenet of the necessity of self-expression for responsible and blameworthy action. Schroeder and Arpaly write:

> It is open to a Whole Self theorist to allow poor integration below some threshold as an excusing condition on moral responsibility [….] An act which is extremely poorly-integrated may be one that, as one might say, is better attributed to circumstances, a drug, or a nefarious neurosurgeon than to the actor, and so be an act for which the actor is not responsible.[37]

One way of putting the main contention of this paper is that this sort of integration is not a necessary condition of responsibility and blameworthiness. I have argued that responsible action can, at times, arise from an extremely "shallow" part of the self: in some cases the only thing expressed by responsible action is the trivial fact that the agent was disposed to act in a certain highly specific way under certain highly specific conditions.

The foregoing points steer us toward a richer understanding of what guides justified reactive feedback. Such an understanding, which can only be reached via consideration of our justified ethical judgments and reactive practices, will leave us with a conception of responsible and potentially blameworthy agency that is much wider than philosophers have imagined, and will save the important everyday concept of the deep self from the encroachment, and potential distortion, of our theories of moral responsibility.

Acknowledgements I have benefited from, and enjoyed, discussing this paper with Susan Wolf, Vida Yao, Thomas Hill, Douglas MacLean, and Ram Neta. I am also grateful for the feedback of an audience at the 2016 Rocky Mountain Ethics Congress, and for the helpful comments. I received there from Daniel Miller. Finally, I am thankful for the careful reading and stylistic wisdom of Pamela Reis.

References

Arpaly, N. (2006). *Merit, meaning, and human bondage: An essay on free will*. Princeton: Princeton University Press.

Arpaly, N., & Schroeder, T. (1999). Praise, blame and the whole self. *Philosophical Studies, 93,* 161–188.

Frankfurt, H. (2003). Freedom of the will and the concept of a person. In G. Watson (Ed.), *Free will*. New York: Oxford University Press.

Nagel, T. (1979). *Moral questions*. Cambridge: Cambridge University Press.

Scanlon, T. (1998). *What we owe to each other*. Cambridge: Belknap Press of Harvard University Press.

Scanlon, T. (2008). *Moral dimensions*. Oxford: Clarendon Press.

Shoemaker, D. (2015). Ecumenical attributability. In R. Clarke, M. McKenna, & A. Smith (Eds.), *The nature of moral responsibility: New essays*. New York: Oxford University Press.

Smith, A. (1976). *The theory of moral sentiments*. Oxford: Clarendon Press.

Smith, A. (2005). Responsibility for attitudes: Activity and passivity in mental life. *Ethics, 115*(2), 236–271.

Smith, A. (2008). Control, responsibility, and moral assessment. *Philosophical Studies, 138*(3), 367–392.

Sripada, C. (2016). Self-expression: A deep self theory of moral responsibility. *Philosophical Studies, 173*(5), 1203–1232.

Walsh shoulders blame for devastating loss (2015). Retrieved from *Espn.com.*

[37] Ibid, p. 174.

Watson, G. (2003). Free agency. In G. Watson (Ed.), *Free will*. New York: Oxford University Press.
Williams, B. (1981). *Moral luck*. Cambridge: Cambridge University Press.
Wolf, S. (1987). Sanity and the metaphysics of responsibility. In F. Schoeman (Ed.), *Responsibility, character, and the emotions*. Cambridge: Cambridge University Press.
Zimmerman, M. (1988). *An essay on moral responsibility*. Totowa: Rowman & Littlefield.