Review of Philosophy and Psychology 1, no. 3 (2010): 319–332

INSANITY, DEEP SELVES, AND MORAL RESPONSIBILITY: THE CASE OF JOJO

David Faraci and David Shoemaker

1. Introduction

To be morally responsible for something is to be eligible for specific sorts of responses to it. Often this will be praise or blame, responses which themselves typically involve various moral emotions (e.g., gratitude or resentment), although actual expression of such emotions is not necessary for responsibility; indeed, you may judge me to be at fault for some action, to be morally blameworthy, without ever actually expressing anything.[1] Furthermore, I may remain eligible for responsibility-responses even when I engage in morally neutral (or merely morally permissible) actions, actions for which neither praise nor blame may be appropriate (e.g., I may be morally responsible for sharpening my pencil). But for me to be eligible for any responsibility-responses with respect to some action, say, that action must at least be properly attributable to me, in some deep and important sense: it must truly be *mine*. That is, the actions for which I am morally responsible must be those that are *expressive of the real me*, expressive of my real self, so to speak.

Or at least this is the thinking behind the *Real Self View* of moral responsibility (the RSV), a label coined by Susan Wolf (1990: Chapter 2) to describe a view originally advocated (in different forms) by Harry Frankfurt, Charles Taylor, and Gary Watson, with a new version introduced more recently by T.M. Scanlon (1998, 2008) and Angela Smith (2005, 2008) (among others). The basic idea has been to identify a subset of an agent's motivating psychological elements as privileged for self-determination and responsibility, such that as long as one's actions are ultimately governed by this subset, they count as one's own and thus render one eligible for responsibility-responses to them.

The different varieties of the RSV depend on what precisely is taken to make some psychological element a member of the privileged set. Frankfurt (1971) famously adopted a hierarchical picture of desires: one's will consists in desires for action about which one may have certain higher-order desires, e.g., desires that those

[1] In a way, this marks the distinction between Gary Watson's "Two Faces of Responsibility," in Watson (2004), pp. 260–288. On the one hand there is, according to Watson, the *aretaic* face of responsibility, in which the relevant responsibility judgments are about moral faults in the agent: his action disclosed something bad about him and his ends, say. On the other hand, there is the *accountability* face, in which being responsible is a matter of being *held* responsible, of being subject to demands of goodwill and being susceptible to certain reactive moral emotions.

first-order desires constitute one's will.[2] The real self is thus located in these higher-order desires. In contrast, Watson located the real self in one's evaluational system, such that free agency (and presumably moral responsibility) consists in one's will depending on or reflecting one's values.[3] And there have been many other variants.[4] But in any event, what these views seem to share in common is the thought that one is morally responsible for some action if and only if it depends ultimately on one's real self.

The RSV, however, has been subject to numerous objections. The worries about it consist mostly in variations on the thought that fulfilling such conditions is actually *unnecessary* for moral responsibility, and there are at least three sorts of cases offered to illustrate the point. First, there are cases of negligence, in which I am responsible for failing to perform some action, where this failure is not obviously governed by my real self.[5] Second, there are cases of whims, willed but unreflective actions that again seem ungoverned by my real self.[6] Third, there are the Huck Finntype inverse *akrasia* cases to which Nomy Arpaly and Timothy Schroeder (1999) have recently drawn attention, cases in which someone may be *praiseworthy* for an action despite its not seeming to flow from that agent's real self.[7]

These are all serious worries, and each of them deserves its own sustained treatment. Here, though, we will be concerned to address a different sort of problem alleged for the view, namely, that meeting the conditions of the RSV is actually *insufficient* for moral responsibility, that one's action could depend on one's real self without one's being morally responsible for the action in question. This is the original objection launched by Susan Wolf against the view in "Sanity and the Metaphysics of Responsibility," (p. 376) and addressing it is essential if one hopes to defend some version of the RSV.[8] So while the objections that RSV is unnecessary for moral responsibility continue to lie in wait, if we can defend the view against Wolf's sufficiency complaint, we will at least have taken a first step towards restoring its viability.

---

[2] At least this was Frankfurt's early take on the matter.

[3] At least this was Watson's early take on the matter. See his "Free Agency," in Watson (2004), pp. 13–32.

[4] See, e.g., Taylor (1976); "Hierarchy, Circularity, and Double Reduction" and "A Desire of One's Own," in Bratman (2007); Stump (1988); Velleman (2002); and Smith (2000). One of our own views is that the right relation consists in one's will depending on psychic elements that ultimately depend on one's nexus of cares. See Shoemaker (2003).

[5] See, e.g., Watson (2004), p. 261.

[6] For this sort of criticism, see Lippert-Rasmussen (2003), esp. pp. 371–373.

[7] See also Arpaly (2003). R. Jay Wallace actually mentions all three of these problems—negligence, whims, and *akrasia*—in Wallace (1994), p. 264.

[8] Wolf mentions the objection again very briefly in Wolf (1990), p. 37.

We begin with a brief discussion of the attractions of the RSV, which ought to provide at least some motivation to rescue it from its objectors.

2. The Attraction

The RSV in its current form developed more or less as a way for compatibilists about responsibility to preserve their view in the face of some difficult real life cases. Early compatibilists (e.g., Hobbes) had a kind of real self view, but at a fairly unsophisticated level. For them, the self was identified with the will (a motivationally efficacious desire), such that if one's will governed one's action, the action was self-determined (and thus, presumably, something for which one was morally responsible). These early compatibilists were attracted to the view because it allowed them to preserve freedom and responsibility in the face of determinism by distinguishing between two types of causation—internal and external—and attaching freedom and responsibility only to the former: as long as I—my will—cause(s) some action, it is free.

This view runs into trouble, however, when we consider cases produced by some *internal* compulsion, e.g., addiction or mania. Actions in such cases are governed by an internal will yet nevertheless seem quite unfree. While some (e.g., libertarians) may be tempted to assert that what is missing in these cases is the robust ability to do otherwise, some compatibilists have reacted instead by pointing to a different lack, namely, genuine *self*-determination. What has gone wrong with the kleptomaniac, for instance, is that the will on which her action depends is not *her* will after all, in an important sense: it is not properly attributable to her in the way necessary for freedom and responsibility. She is instead alienated from it. And so what is required to render a will one's own is that it depend on one's *deepest* or *most true* self, whatever that may be. Consequently, contemporary RSV theorists typically incorporate two (or more) levels (e.g., of desires) or motivational systems, such that the real self is located at the level or in the system that can reflect on and govern its will (or "superficial self").

This view can be interpreted as thwarting not only the threat of determinism but also a more general skepticism about responsibility, a skepticism based on the intuition that one is responsible for one's actions only to the extent one is responsible for the self that performs those actions. But insofar as this intuition seems to presuppose an impossible kind of self-creation, responsibility for one's actions seems impossible.[9] Given that the RSV's adherents run a key distinction between the acting, superficial self, and the responsible, real self, however, they can account for the intuition while detaching it from the worry about self-creation: if we take up the

[9] One might see traces of this objection in Strawson (2003).

phenomenological perspective, we will see that it is enough for us that we can reflect upon, revise, and generally govern our superficial selves, given that this ability provides us with "all the freedom it is possible to desire or to conceive." (Frankfurt 1971, p. 16) Responsibility for our superficial selves, in the sense of their being under our control, is all we require for responsibility generally.[10]

Independently of its developmental role in the history of compatibilism, however, the appeal of the RSV is rather obvious. Surely for an action to be my own, to belong to me for purposes of moral assessment, it must flow in some way from my real self. If it did not have its source in the true me, it is difficult to see moral assessment *of* me for that action as having much purchase. To judge me blameworthy for some action, for example, is to find fault *with me,* and it is hard to see this assessment as being coherently directed to anything other than my real self. But if such assessment is directed to my real self, that self must have some governance over the actions attributed thereto; otherwise, such assessment would be deeply unfair. Furthermore, if that governance relation is in place, it is rather difficult to think of anything missing needed to render the action flowing from such governance one's own. When I—my real self—govern the action (in the relevant sense), I have in effect "mixed my labor" with those bodily movements in the only way in which, it seems, they can sensibly be rendered my property.

The motivation for the RSV, then, is relatively simple. One might worry, though, that it is *too* simple, that it fails to account, for instance, for complications in the formative circumstances of the governing relation in question. Indeed, this is precisely where Wolf attacks.

3. The Objection

In her article on the metaphysics of responsibility, Wolf discusses a version of the RSV she actually calls the "deep self view" (DSV), according to which one is responsible for some action A just in case (1) A is governed by one's will, and (2) one's will is governed by one's deep self (where this is located either in one's highest-order desires or one's evaluative judgments). Her objection to this theory is based on her famous case of JoJo:

> JoJo is the favorite son of Jo the First, an evil and sadistic dictator of a small, undeveloped country. Because of his father's special feelings for the boy, JoJo is given a special education and is allowed to accompany his father and observe his daily routine. In light of this treatment, it is not surprising that little JoJo takes his father as a role model and develops values very much like Dad's. As an adult, he does

[10] Cf., Wolf (2003), pp. 376–379

> many of the same sorts of things his father did, including sending people to prison or to death or to torture chambers on the basis of whim. He is not *coerced* to do these things, he acts according to his own desires. Moreover, these are desires he wholly *wants* to have. When he steps back and asks, "Do I really want to be this sort of person?" his answer is resoundingly "Yes," for this way of life expresses a crazy sort of power that forms part of his deepest ideal. (Wolf 2003, p. 379)

Here JoJo thinks his actions involving torture and brutality are *morally* required, expressions of values to which he is deeply committed. The contrary moral demands we would make of him would be met with a kind of puzzled disdain. He would fail to see our demands as demands of morality; they would simply find no purchase in the value system embedded in him since childhood (a heritage and upbringing he was "powerless to control" (p. 379)). Nevertheless, as Wolf implies, the structure of his will is such that he meets both of the DSV's conditions for moral responsibility: his actions are governed by his will; and his will is governed by his deepest self. But, as Wolf maintains, given the formative circumstances of his will, it is in fact "dubious at best that he should be regarded as responsible for what he does," for it "is unclear whether anyone with a childhood such as his could have developed into anything but the twisted and perverse sort of person that he has become." (pp. 379–380)

Of course, one might well suggest the source of JoJo's lack of responsibility stems from the point discussed earlier, namely, the fact that his deepest self was not up to him insofar as he had no governance over its creation. What Wolf does instead is remind us that not everything that matters for responsibility has to be about power and control; it may be that part of what matters is that we simply *be* a certain way, regardless of whether or not we had any control over being or becoming that way. And we find support for this thought in considerations of legal responsibility, a key condition for which is the *sanity* of the accused. Adopting this condition for moral responsibility, then, we can see that even though JoJo's will is structured properly on the DSV, he is not yet eligible for responsibility because he is normatively *insane*: he lacks "the ability to know the difference between right and wrong…." (p. 382) That is, JoJo's values have not been controlled properly by the way the world is, and given his formative circumstances he could not help but have been mistaken about those values. To recognize this fact is thus to recognize that JoJo lacks a key condition for moral responsibility. On Wolf's own view, then, one is eligible for responsibility only if (a) one can govern one's actions with one's desires, (b) one can govern one's desires with one's deep self, and (c) *one's deep self is sane.* (p. 382) This view—the Sane Deep

Self View—allows us to maintain the initial plausibility of the DSV while also being able to account for the case of JoJo.

Wolf goes on to suggest that her account reveals that certain other historical immoralists, e.g., Nazis, slaveholders, and male chauvinists, were not fully responsible either, precisely because they were also (partially) normatively insane. These sorts of folks are victims of either "deprived childhoods" or "misguided societies," and their actions "are governed by mistaken conceptions of value that the agents in question cannot help but have." (p. 383) It is the "mistaken" aspects of their value conceptions that matter here: we have all been raised with conceptions of value we cannot help but have; but that alone does not incapacitate us from responsibility. It must instead be only when our value system does not allow us "normatively to recognize and appreciate the world for what it is" that we are insane and thus not responsible. (p. 383) Wolf's objection to the DSV, then, is that it is incomplete: the type of self-expression both necessary and sufficient for moral responsibility, she avers, is the self-expression of a *sane* self. If the sanity condition is included, then (and only then) could something like the DSV (and thus the RSV) be true.

4. The Experiment

Wolf takes it as a fundamental datum that our pre-theoretic intuitions converge on the judgment that JoJo is not a responsible agent. But do they? Whenever we introduced the case to students, they always needed considerable coaching to come to the conclusion Wolf wants. They resisted the idea that JoJo is not responsible primarily because they found it extremely hard to believe that JoJo would not be able to recognize that torture is wrong. In response, various features of the case would have to be stressed or exaggerated, e.g., the isolation in which JoJo grew up, in a "small, undeveloped country"—an island, it was proffered, with no communication links to the outside world, with heavily propagandized internal media, etc. Eventually the students would reluctantly agree to the "intuition," but at a price: the case now seemed quite precious. JoJo was now taken to live in an airtight vacuum, cut off from the world as we know it, and he was being rescued from responsibility, if at all, by a kind of forced and surreal *ignorance*. When the case was brought back into the real world, focused on someone like Uday Hussein, son of Saddam, the intuition that he was responsible seemed to return full-force.

To this point all we had were anecdotes and speculation, however, so it was decided that it would be worthwhile to gather data and find out just what people's pre-theoretic intuitions on the case really were. Our hypothesis was that JoJo would likely still be thought to be blameworthy, albeit perhaps less blameworthy than someone like Uday Hussein, and that the degrees of blameworthiness would increase

in correspondence to his exposure to relevant moral alternatives.[11] In testing this hypothesis, we wrote up three scenarios, one a control, one Wolf's original JoJo scenario, and one in which Wolf's JoJo is exposed to moral alternatives. The subjects of the survey then each received one of these scenarios and were asked after reading it to "circle the number that best represents the degree to which you think JoJo is *blameworthy* for his actions (sending people to prison/death/torture chambers)" on a scale of 1 to 7, with "1" being not at all blameworthy, "4" being somewhat blameworthy, and "7" being completely blameworthy.[12] The first—control—scenario was what we took to be Jo the First's story (although we used the name "JoJo" to prevent any worries from arising about different names being deployed in different scenarios):

> *JoJo1 Scenario:* JoJo is an evil and sadistic dictator of a small, undeveloped country. JoJo does many things as a dictator, including sending people to prison or to death or to torture chambers on the basis of whim. He is not coerced to do these things. When he steps back and asks, "Do I really want to be this sort of person?" his answer is resoundingly "Yes," for this way of life reflects his deepest values and ideals.

Answers here would help establish a baseline for responses to JoJo's particular brand of nastiness in the absence of any specified upbringing or exposure to alternatives.

The second scenario was almost identical to Wolf's own:

> *JoJo2 Scenario:* JoJo is the favorite son of Jo the First, an evil and sadistic dictator of a small, undeveloped country, entirely cut off from the outside world. Because of his father's special feelings for the boy, JoJo is given a special education and is allowed to accompany his father and observe his daily routine. In light of this treatment, little JoJo takes his father as a role model and develops values very much like Dad's. As an adult, JoJo does many of the same sorts of things his

[11] The language used here matters. In Wolf's original scenario, the relevant assessment is that JoJo is not to be regarded as "responsible" for what he does (and of course the title of the article itself declares it as being about "responsibility"). "Moral responsibility" is something of a term of art in philosophy, however, and it's not at all clear what the folk in general take it to mean. It could, after all, be taken to mean anything from "causally responsible" to "having a moral duty" to "having a role-based duty" to "being open to moral appraisal" to "living up to moral expectations" (according to which JoJo could be thought to be quite *ir*responsible). We thus decided to test intuitions using the term "blameworthy," as this is less of a term of art, is likely clearer, and is still in accordance with what Wolf had in mind, it seems.

[12] All were students at Bowling Green State University.

> father did, including sending people to prison or to death or to torture chambers on the basis of whim. He is not coerced to do these things. When he steps back and asks, "Do I really want to be this sort of person?" his answer is resoundingly "Yes," for this way of life reflects his deepest values and ideals.

Changes here from Wolf's original were minimal, but are worth explaining. In the first sentence, we added the phrase "entirely cut off from the outside world" in order to emphasize the isolation of JoJo's upbringing, thus leaving it open that JoJo's wholehearted adoption of his father's values might be due to simple lack of access to moral alternatives. In the third sentence, we deleted Wolf's commentary that "it is not surprising" that JoJo develops values very much like Dad's, as we did not want to prejudice respondents one way or the other as to the actual degree of surprisingness attached to such a development. In the final two sentences, we eliminated references both to the technical machinery Wolf is criticizing (e.g., references to hierarchical desire systems) and to her assertion that JoJo's way of life expresses a "crazy sort of power"—obviously, an attempt to determine in a neutral fashion whether or not JoJo is morally insane could be derailed by any talk beforehand of his "craziness."

The final scenario changed only JoJo's exposure to alternative value systems:

> *JoJo3 Scenario:* JoJo is the favorite son of Jo the First, an evil and sadistic dictator of a small, undeveloped country, entirely cut off from the outside world. Because of his father's special feelings for the boy, JoJo is given a special education and is allowed to accompany his father and observe his daily routine. In light of this treatment, little JoJo takes his father as a role model and develops values very much like Dad's. When he turns 21, JoJo is sent to live in a developed country for a year, and there he becomes aware that other leaders treat their subjects with respect and goodwill because they value the lives and well-being of their subjects. Nevertheless, when he returns to lead his country, JoJo does the same sorts of things his father did, including sending people to prison or to death or to torture chambers on the basis of whim. He is not coerced to do these things. When he steps back and asks, "Do I really want to be this sort of person?" his answer is resoundingly "Yes," for this way of life reflects his deepest values and ideals.

5. The Results

As expected, there was a difference in assessments of blameworthiness between the first two cases, but not at all of the sort Wolf envisioned. On average, subjects did

judge JoJo1 (the control) quite sternly: the mean was 5.8 (on a scale of 1–7), with exactly half of the respondents assigning a solid 7 to him. In the JoJo2 case, however—essentially Wolf's original JoJo case—the mean was 4.77. This difference between the average assessments of JoJo1 and JoJo2 was found to be statistically significant.[13] So, while subjects did deem JoJo2 to be less blameworthy than JoJo1, they still judged him to be more than somewhat blameworthy. (The mode score was actually a 6, and the median reply was a 5).

What of the responses to JoJo3, the one seemingly aware of moral alternatives? The responses here were unexpected: the mean was 4.93, but the mode and median scores were both 5, trends suggesting that subjects found that someone with JoJo's background is less blameworthy than Jo the First, even if he has been exposed to moral alternatives. The differences in mean between JoJo2 and JoJo3, however, were not statistically significant.

6. The Philosophical Discussion

What do these results mean for Wolf's claim about the incompleteness of the DSV (and thus the RSV)? First off, they seem to render the charge unmotivated. Wolf's case is built explicitly and entirely on our pretheoretic intuitions in the JoJo case, and given that these do not come close to converging on the view that JoJo is not blameworthy, we do not yet have a good reason to believe that meeting the conditions of the DSV may be insufficient for moral responsibility. Indeed, people clearly attribute the nasty actions to JoJo2, and insofar as the action depends on his will, which itself depends on his deep self, the DSV provides a straightforward story about why he is responsible and blameworthy: he is actively governing his own actions.

Nevertheless, there are some complications here. First, our hypothesis (based on anecdotal classroom experience) was that the explanation for any difference between JoJo1 and JoJo2 would be his lack of exposure to moral alternatives—that is, his *ignorance,* not his insanity. The subjects' responses to the JoJo3 case seem to suggest otherwise, however. Judgments of his blameworthiness were essentially no different from JoJo2's, so it seems that consideration of JoJo's unfortunate formative circumstances dominates his level of exposure to moral alternatives: these circumstances render him less blameworthy (according to our subjects) than the

[13] For each survey, N=30. The overall ANOVA demonstrated a statistically significant difference between average responses for all three JoJo cases ($F=4.318$, $p=.016$). Bonferroni post-hoc comparisons between responses for JoJo1 and JoJo2 demonstrated a significant tendency to judge JoJo2 less responsible ($p=.023$). Comparisons between JoJo1 and JoJo3 demonstrated a trend towards judging JoJo3 less responsible ($p=.072$). Comparisons between JoJo2 and JoJo3 demonstrated no significant difference in responses ($p=1$).

control case seemingly *regardless* of his moral ignorance. This result genuinely surprised us.

Beyond being a surprise, though, it also opens the door for a Wolfian response to our interpretation of subjects' assessments of JoJo2. Admittedly, she might say, subjects do still find the original JoJo to be pretty blameworthy. But he is most definitely *less* blameworthy than JoJo1, where the only difference between them is the statement of JoJo2's unfortunate formative circumstances. If inclusion of exposure to moral alternatives does nothing to alter that assessment, then it looks as if that exposure could do nothing to alter JoJo's commitment to Daddy's ideals. But if this is the case, then it seems like JoJo really does lack the ability to adopt the right value system even with exposure to it—he is normatively insane. Consequently, the DSV remains incomplete without a sanity condition.

This is much too quick, however. For one thing, we cannot forget that JoJo2 was deemed to be quite blameworthy (indeed, *no one* thought he was not blameworthy at all). So it is not at all clear that his being (slightly) less blameworthy than JoJo1 warrants judging the DSV as it stands to be insufficient. Indeed, one might think that he is as blameworthy as he is precisely because he meets the conditions of the DSV, just not as *fully* as does JoJo1. Perhaps subjects think that JoJo2's actions are not as attributable to his deep self as JoJo1's, given the former's upbringing, or perhaps subjects think that the *depth* of JoJo2's deep self is somewhat limited. But in any event, there may be scalar resources within the DSV itself that could explain both the significant degrees of blameworthiness attached to both JoJos as well as the disparity between them. (More on this point later.)

Indeed, it may still be the case that moral ignorance is playing a key role in the disparity between JoJo1 and JoJo2. After all, people could well be judging that mere exposure to moral alternatives is insufficient to dispel moral ignorance, especially for those who have been as thoroughly indoctrinated as JoJo. It is a common phenomenon, after all, for people to exhibit confirmation bias—to be unwittingly selective in what evidence they gather and deploy.[14] Subjects may well be thinking that mere exposure to moral alternatives is itself insufficient to undermine JoJo3's deep-seated moral ignorance, believing that JoJo may be unconsciously disposed to ignore alternatives to his deep-seated values.

JoJo2 and JoJo3 may be thought to be closely aligned in this respect precisely because of the particularly insidious type of ignorance they have in common, a type we believe is different in *kind* from the moral ignorance people usually experience. For most of the rest of us, moral ignorance consists in our lacking knowledge that some particular action is generally construed as an expression of ill will, where such expressions are what we typically believe make the action wrong. In other words, we

[14] See, for one discussion, Nickerson (1998).

are aware of which moral properties supervene on which act-types, but we may lack knowledge of which act-tokens fall under those act-types. This matters to us, then, because insofar as some act-token is perceived to express ill will, it is something we will hasten to correct (by apologizing or explaining, say): we (typically) want to express goodwill in our actions; and if they are perceived otherwise, we want it known that we did not in fact intend ill will at all.

Now when (either) JoJo inflicts his injuries on the peasants he actually intends to do so: this is the only way such peons will learn, he thinks, or perhaps it expresses a power suitable to his station. We might thus say that JoJo intentionally expresses *ill* will to them. What is he ignorant of, then? He does not know that expressions of *ill will* are wrong. That is, his is a more fundamental ignorance than ours: he is unaware of which moral properties supervene on which act-types. Indeed, he thinks that what he is doing is morally right, that he is following squarely in the footsteps of his "admirable" and "morally good" father. This is a kind of ignorance, though, that mere exposure to the relevant moral alternative—the basic demand that expressions of ill will are wrong—may not be sufficient to displace. Further, if this analysis is on the right track, then while JoJo's actions depend on his motives, and his motives depend on his deep self, he may not be seen as fully violating the basic demand for goodwill unless he is first aware *that* it is a demand and *that* he is violating it. Neither condition is fulfilled in this case and this is one reason, we suggest, that subjects' judgments of his blameworthiness might be dampened.

The difficulty in disentangling Wolf's analysis of the case from our preferred ignorance-based analysis stems from the fact that normative insanity entails normative ignorance: If I lack the *ability* to know that X is wrong, then clearly I *do not* know that X is wrong. But how can we determine that the latter has its source in an incapacity, rather than in some other sort of (mere) deficiency? Wolf claims to follow the M'Naughton Rule in legal contexts, which holds that a person is sane "if (1) he knows what he is doing and (2) he knows that what he is doing is, as the case may be, right or wrong." (Wolf 2003, p. 381) Wolf's gloss on the second condition is that the sane agent has "the *ability* to know the difference between right and wrong…." (p. 382, our emphasis) But it is not clear that JoJo lacks the stated ability at all. For one thing, if "rightness" and "wrongness" are roughly construed as "to be doneness" and "not to be doneness," then JoJo would seem to have precisely the same level of ability in this regard as most of the rest of us. What JoJo lacks is knowledge of what *is* right and what *is* wrong, not the difference between them. In addition, it remains unclear that JoJo's lack of knowledge stems from a disability (the disability essential to insanity): just because he does not know that expressions of ill will are immoral, that does not yet mean that he lacks the ability to know it. Until this gap between moral ignorance and moral insanity is closed, the DSV's incorporation of a sanity condition seems unmotivated.

Nevertheless, there is still a puzzle here. If ignorance were a genuine excuse, then perhaps one ought to think, right along with Wolf, that JoJo's actions really should not be actions for which he is responsible *at all.* In ordinary cases, for example, once we find out about another's ignorance in performing some injurious action, all blame is typically withdrawn. But in the experimental scenario, JoJo2 has clearly been deemed blameworthy *to a significant extent,* just as has JoJo3. One problem here is that the DSV as it stands (and as mentioned earlier) does not seem sensitive to scalar assessments of responsibility. Instead, it holds that one is responsible for one's actions just in case one's actions depend on one's will and one's will depends ultimately on one's deep self, where this is an all-or-nothing assessment: if one fulfills the conditions, one is responsible, and if one does not, one is not. Thus, given that JoJo2 and JoJo3 fail to meet the first condition, they should not (on this view) be eligible for responsibility assessments at all.

We speculate, however, that what may mitigate the full-blown excusal of JoJo2 is the belief that his ignorance itself is rather culpable, that even though he did not in fact know better (and his ignorance is deep-seated), he *should have,* where this means there were plenty of opportunities for him to infer that expressions of ill will were wrong, if only he had paid closer attention or been sufficiently sensitive.[15] Indeed, subjects may well have thought that JoJo3's exposure to the moral alternative was *redundant*: he had already had sufficient exposure to render his ignorance culpable. Nevertheless, because of the deep-seated nature of the ignorance, subjects could be cutting both JoJo2 and JoJo3 roughly equal slack, given the degree of difficulty attached to their actually succeeding in identifying and eliminating their zones of ignorance.[16] People thus could be thinking of both JoJo2 and JoJo3 (and those like them) in the following way: while it would be difficult for them to come to see

[15] This is closely akin to finding fault with someone for her *negligence.* Negligence is notoriously difficult for theories of responsibility to handle adequately, in part because there doesn't seem to be any particular *action* for which the agent is blameworthy; rather, it's an *absence* of an action that garners fault-finding, e.g., someone's failure to salt the ice off her sidewalks before hosting a big party (where someone falls on the ice), or someone's being so engrossed in a conversation that he fails to look both ways and plows into another car. (For a brief discussion of the latter example, see Gary Watson, "Skepticism About Weakness of Will," in Watson (2004), pp. 52–53.) We have no intention of solving this difficult problem here; rather, we hope to show how the scalar dimension of some assessments of responsibility may derive from considerations in this neighborhood.

[16] It is also worth pointing out here that we are not entirely convinced that the lack of difference in assessments of JoJo2 and JoJo3 was not just due to the fact that our sample size was too small. This is, at the very least, a survey worth running again, perhaps to see if there is a difference between responses to a JoJo who has been exposed to expressions of good will and a JoJo whose exposure has actually penetrated his ignorance.

the attraction of the basic moral demand, given their upbringing, it is nevertheless not unreasonable for us to expect them to do so. The fact that they do not know, therefore, while ordinarily exculpatory, is not sufficient to get them off the hook here. They are thus still blameworthy for their actions to a degree corresponding inversely to the degree of difficulty attached to their incorporation of the relevant moral awareness: the more difficult, the less blameworthy.

7. The Conclusions

The story told by the advocate of DSV in response to the JoJo case is rather complex, but it does not at all require either an admission of the insufficiency of the account or any reference to sanity or insanity. Instead, one could still conclude as follows:

1. The experimental results strongly suggest that Wolf's assumptions about our pretheoretic intuitions in the JoJo case are wrong. Rather than believing that JoJo is, given his upbringing, not responsible, most seem to believe that JoJo is blameworthy (awhich entails his responsibility)—just not as blameworthy as he might have been. This evidence alone seems to undermine Wolf's motivation for saying the DSV is insufficient, and it also obviates the need to add a sanity condition. (And insofar as the DSV is just a specific version of the Real Self View, her charge has no purchase for that more general view either.)

2. Suppose, though, that the evidence is a fluke, or that Wolf really wants to target more sophisticated intuitions, and that a JoJo-like scenario actually does cause people to judge JoJo not to be blameworthy. Would *this* turn of events motivate a need for a sanity condition to supplement the DSV? No. What may still ground such intuitions could be assumptions about JoJo's critical ignorance of the moral alternatives to the wrongness of his expressions of ill will (where this may now be viewed as non-culpable ignorance): failure to know does not necessarily imply inability to know. But ignorance-as-undermining-responsibility may easily be accounted for by the DSV, insofar as ignorance of what some piece of behavior actually consists in *qua* action (under the proper description) disrupts the dependence relation between one's superficial self/will and one's action: one would not be intending to perform the specific action for which one was being held responsible.

3. Nevertheless, given the evidence as it stands, JoJo2 and JoJo3 *are* thought to be blameworthy to a significant degree. How might we explain this? According to the DSV, if any JoJo is deemed blameworthy, it must be in virtue of the fact that his immoral action depended on his will, which itself depended on his deep self. So far, so good: both *are* deemed blameworthy, and both in fact do meet the conditions of the DSV. The puzzle, though, is why they are thought *less*

blameworthy than the control, JoJo1. How might this difference in degree be accounted for on the DSV? One answer that we have suggested points to a tension in our intuitions. On the one hand, there is the pull towards judgments of non-responsibility (or at least non-blameworthiness) that could be explained by thinking of the JoJos as morally ignorant, of failing to see the basic moral demand as a genuine alternative for them. On the other hand, there is the pull towards blameworthiness suggested by the thought that, even if they did not know of the alternatives, they *should have* known, such that their ignorance is morally culpable (to some extent).

One possible problem with both aspects of this analysis, though, is that there was not much of a difference between reactions to JoJo2 and JoJo3, where the latter had actually been exposed to the relevant alternatives. What we have suggested in reply, though, is that subjects might still be viewing JoJo3 as just as ignorant as JoJo2, that given the type and the deep-seated nature of his ignorance, his being exposed to alternatives in this particular way might make no discernible difference to him. Of course, this view seems to bring out an important tension: if his ignorance is impervious to this sort of exposure, why think it is sufficiently culpable to support the judgment of significant blameworthiness nevertheless? But this is only a tension if we think that JoJo2 and JoJo3 had no *other* opportunities to incorporate the relevant moral awareness, and if we do not, it would be easy to maintain that JoJo3's brand of exposure was just redundant by that point.[17]

4. An important adjustment needs to be made to the DSV for it to be able to account precisely for these results. As it stands, it does not reflect a scalar component, but subjects do make assessments of varying degrees of blameworthiness if given the opportunity. We have suggested that differences in degrees of assessment may correspond inversely to the degree of difficulty in the target's coming to moral recognition. On this explanation, then, an advocate of the DSV could claim that the degrees of blameworthiness track the degrees of *attributability of actions*: actions are more or less attributable to agents in these sorts of cases depending on the degree of difficulty they are judged to have in recognizing various features of their actions about which they remain ignorant. But regardless of the specific explanation of the data, a scalar dimension may be incorporated into the DSV in the following way: one is

[17] It is also worth pointing out here that we are not entirely convinced that the lack of difference in assessments of JoJo2 and JoJo3 was not just due to the fact that our sample size was too small. This is, at the very least, a survey worth running again, perhaps to see if there is a difference between responses to a JoJo who has been exposed to expressions of good will and a JoJo whose exposure has actually penetrated his ignorance

morally responsible only for *one's own* actions, of course, but ownership may come in degrees, and as a result so may blameworthiness.

5. Wolf maintains that what JoJo lacks is not awareness of alternatives but a certain *capacity*, namely the capacity to recognize the difference between right and wrong, or, more plausibly, the capacity to recognize the *facts* about right and wrong, a capacity akin to the cognitive capacity to recognize certain (non-moral) facts about the way the world really is. So if you cannot recognize that torturing babies, say, is immoral, then there is something wrong with you, independently of your lack of exposure to the basic demand for goodwill or your genuine ignorance about alternative conceptions of value. This view seems too strongly realist, however, requiring as it does moral facts as somehow part of the actual fabric of the world, capable of being recognized by sane people in the same sort of way facts about physical objects are capable of being recognized. While some folks adhere to a view like this, it is clearly at odds with views denying moral facts altogether, as well as with constructivist views like contractualism, according to which the moral facts are constructed via reasonable agreement amongst similarly motivated (and reasonable) people.[18] On these views, failing to "recognize" that torturing babies is immoral does not, in and of itself, indicate a normative incapacity; it may simply, as we believe in the case of JoJo, indicate an especially high degree of isolation and lack of genuine human interaction in upbringing that lends itself particularly well to dangerous ignorance. It is not necessarily that JoJo *cannot* see the basic demand; perhaps, instead, he simply *does* not. So while he may be viewed as less blameworthy for his current actions, it is far too quick to claim that he is *exempted* from the world of moral agents. To call JoJo insane, then, is to throw up our hands in a way the case does not yet warrant.

6. The DSV may thus be defended against Wolf's charge that it is insufficient to account for moral responsibility (although surely more studies are needed to explore the relation between ignorance and insanity). How the Real Self View, its more general parent, fares with respect to the charge that it fails to provide the *necessary* conditions for responsibility, however, is a story for another day.

Acknowledgements

The authors would like to extend their deepest thanks to Leisha Colyn for her invaluable help and advice with the experimental aspects of this paper. We would also like to thank two referees for the special issue of the European Journal of

[18] See Scanlon (1998). See also Derek Parfit's forthcoming book *On What Matters.*

Philosophy, one of whom revealed himself to us as Eddy Nahmias. Finally, a special debt of gratitude goes to Joshua Knobe, whose ongoing encouragement and enthusiasm for this project was truly the spark of its ultimate completion.

References

Arpaly, N. 2003. *Unprincipled virtue: An inquiry into moral agency.* Oxford: Oxford University Press.

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

Bratman, M. 2007. *Structures of agency.* Oxford: Oxford University Press.

Frankfurt, H. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* LXVIII: 5–20.

Lippert-Rasmussen, K. 2003. Identification and responsibility. *Ethical Theory and Moral Practice* 6: 349– 376.

Nickerson, R.S. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 2: 175–220.

Scanlon, T.M. 1998. *What we owe to each other.* Cambridge: Harvard University Press.

Scanlon, T.M. 2008. *Moral dimensions.* Cambridge: Belknap Press of Harvard University Press.

Shoemaker, D.W. 2003. Caring, identification, and agency. *Ethics* 114: 88–118.

Smith, A. 2000. Identification and responsibility. In *Moral responsibility and ontology,* ed. T. van den Beld, 233–246. The Netherlands: Kluwer Academic Publishers.

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

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

Strawson, G. 2003. The impossibility of moral responsibility. In *Free will,* 2nd ed, ed. G. Watson, 212– 228. Oxford: Oxford University Press.

Stump, E. 1988. Sanctification, hardening of the heart, and Frankfurt's concept of free will. *Journal of Philosophy* LXXXV: 395–420.

Taylor, C. 1976. Responsibility for self. In *The identities of persons,* ed. A.O. Rorty, 281–299. Berkeley: University of California Press.

Velleman, J.D. 2002. Identification and identity. In *Contours of agency,* ed. S. Buss and L. Overton, 91– 123. Cambridge: MIT.

Wallace, R.J. 1994. *Responsibility and the moral sentiments.* Cambridge: Harvard University Press.

Watson, G. 2004. *Agency and answerability.* Oxford: Oxford University Press.

Wolf, S. 1990. *Freedom within reason.* Oxford: Oxford University Press.
Wolf, S. 2003. Sanity and the metaphysics of responsibility. In *Free will*, 2nd ed, ed. G. Watson, 372–387. Oxford: Oxford University Press.