Macalester Journal of Philosophy

Volume 11
Issue 1 *Spring 2002*

Article 11

3-10-2011

Wills, Persons, and Moral Responsibility, or, No, Virginia, There Really is not a Sanity Clause

Marc Johansen
Macalester College

Wills, Persons, and Moral Responsibility, or, No, Virginia, There Really is not a Sanity Clause

Marc Johansen

It is widely held that problems of personal identity are deeply related to problems of responsibility. While some philosophers such as Derek Parfit may deny such a relation, it is in the very least true that we possess the intuitive belief that the domain of moral responsibility is populated solely by persons. As such, the metaphysical status of persons would appear to be deeply relevant to questions of moral responsibility. This relationship puts the issue of determinism in a special light. For if we as persons are not just actors but products, we must sharply question the sense in which we may be responsible for actions which, strictly speaking, have their causal origin outside of us.

In this essay I will examine a compatibilist response to this question which Susan Wolf refers to as the "deep self view." This position is best characterized in the work of Harry Frankfurt (e.g., "Freedom of the Will and the Concept of a Person," "Identification and Externality," "Identification and Wholeheartedness," etc.) and Gary Watson in his essay entitled "Free Agency". While there are subtle and important differences between the positions Frankfurt and Watson advocate, the concepts at the core their work are essentially equivalent, such that they can be seen as offering variations on the same underlying

---

[70] Dostoevsky: *Notes From Underground*, pp. 192.

position.[71] It is my aim here to demonstrate that the strategy presented by the deep self view is rooted in a highly suspect metaphysics and as such is not merely incomplete but fundamentally flawed.

The central tenet of the deep self view is that our basic motivational desires are controlled by a deeper, underlying will. In Frankfurt's work this idea is expressed in terms of a hierarchy of desires; to Watson the distinction is between basic desires and those that are values.[72] Behind both these positions, however, lies the intuition that simple intentionality and purposeful behavior is insufficient grounds for moral responsibility. Responsibility, rather, hangs on the fact that our basic desires, which may be shaped by causal forces external to us, can be evaluated by a deeper level of self in which the self proper - the person - is located. Through this evaluative process, persons may identify with those desires they want to consider part of themselves. Such desires, regardless of their origin, become internalized, and therefore part of the self's arena of responsibility. As such, whatever factors may influence basic motivational desires, persons may be held responsible for actions springing from those desires they have chosen to identify with.

---

[71] Unless explicitly noted otherwise, the deep self view I will refer to from here forward is this common perspective which stands behind both their work.

[72] This difference in characterization of the deep self is in fact the primary distinction between the Frankfurt and Watson views of the deep self. While Frankfurt is content to regard higher order desires as otherwise ordinary desires directed towards those that are more basic (e.g., I don't want to want that cookie), Watson insists that deeper values are of qualitatively different type than shallow desires.

This structural account of persons provides a strong account of a number of important moral intuitions regarding responsibility and external influence. Most relevant to this discussion are the cases of overdetermination and the inability to bring shallow desires in line with their deeper counterparts. Frankfurt illustrates these cases through the study of an addict. In either case, the addict possesses an overwhelming desire to take some drug. The force of this desire is such that the addict is compelled to take the drug irregardless of what he wants to desire. The actual outcome of either case, then, is the same. However, in the latter case, the addict is unwilling - he does not want to feel compelled to take the drug - while in the former, the addict is quite happy with his compulsion. Frankfurt points out that in the case of the unwilling addict we do not hold him responsible for the fact that he takes the drug. In this particular instance, the motivating force behind the act is external to the person; despite his wishes to the contrary, he is compelled by his basic desire for the drug.

The case of overdetermination is less intuitive. While the willing addict has no more control of his final action then his unwilling counterpart, the deep self view holds the willing addict responsible for his behavior. This is due to the fact that while external forces may compel the actions of the willing addict, his internal desires, those he has identified with, push him towards the same action. For even in the absence of this addiction we have strong reason to believe that he would have behaved in the same fashion.

It would seem then that the deep self view provides a sound account of at least two primary aspects of moral responsibility. While

these two cases are not responsibility's whole story, the deep self model appears to provide a plausible ground for a concept of moral responsibility. Consider, however, the following situation:

> 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 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. (Wolf, 54)

It seems questionable at best that we can hold JoJo to be morally responsible for his actions in any sort of deep way. The values he came to hold were a direct result of his upbringing, to which he cannot be held accountable. His character was shaped, though clearly not against his will, in such a way that it seems highly unlikely that he would grow to hold any other set of values. Moreover, there is nothing special about JoJo in this case. It seems unlikely that anyone placed in the same relationship as JoJo to the historical, causal forces that were involved in his upbringing would have fared better.

However, while we may cringe at the thought of possessing such a character as does JoJo, it cannot be said that he feels any such regret for his fate. Quite the contrary, he would have it no other way. The values that JoJo acquired from his upbringing are ones that he

wholeheartedly approves of and feels no compulsion to change. They were willingly acquired and have been identified with on a very deep level.

While the deep self view may provide a sound theoretical framework for overdetermination and runaway basic desires, it flounders in cases such as JoJo's. These problem cases are by no means special either logically or empirically; they illuminate flaws that lie at the heart of the deep self theory. Cases such as JoJo's make clear that the will, however deep it may be located, is not singly capable of providing persons with a sufficient degree of responsibility over their character.

The problem for the deep self view is a structural one. As noted previously, the deep self view attempts to attain some sense of control over the self, and therefore responsibility for actions performed by that self, by shifting the "real" self from the level of direct, motivational desires to a valuational level capable of selecting those desires that it chooses to identify with. As the case of JoJo demonstrates, however, it's not entirely clear how such a shift removes the self from the path of the causal, historical forces which raised the issue in the first place. It would seem that this move only changes the level at which we present the problem rather than alter the situation in any significant manner. The concern is now whether we are in some way responsible for the values that we hold.

The clear, but equally problematic, solution to this dilemma is to hold that this second level of desire is in fact governed by a still deeper layer. Frankfurt suggests just that, stating in "Freedom of the Will and the Concept of a Person", that "There is no theoretical limit to the length

of the series of desires of higher and higher orders" (446). While positing still deeper levels of will in which we might bury the self may produce a temporary solution to this problem, it is unclear how such a response can avoid an infinite regress. However many levels of depth are invoked to ward off the boogeyman of nonresponsibility, in all cases there will be some level such that there are none beyond it and we are now forced to consider how this final level is controlled. If this is the case, the deep self is no more free than are one's most basic desires. Any reason that might be devised for the control of this final order of desire could just as easily be applied to the first.

Frankfurt insists, however, that an infinite regress can be avoided through the principled termination of a series of "deeper" desires. He claims that:

> when a person identifies himself decisively with one of his first-order desires, this commitment "resounds" through the potentially endless array of higher orders.... The decisiveness of the commitment he has made means that he has decided that no further question about his second-order volition, at any higher order, remains to be asked (Frankfurt, 15, italics his).

But this answer, as stated, is entirely insufficient to provide an escape from the infinite regress that seems inherent to his position. The concepts he cites in the above passage are far too vague to do the work necessary to halt the regress faced by the deep self view. It is unclear what is intended by the term "decisively" or how such a commitment "resounds" through the levels of desire, much less how such concepts

would accomplish what Frankfurt intends for them. As it stands, Gary Watson quite correctly notes in his critique of Frankfurt's view of the will that, "It is unhelpful to answer that one makes a 'decisive commitment,' where this just means that an interminable ascent to higher orders is not going to be permitted. This is arbitrary." (218).

Frankfurt does go on to clarify this point in other writings, but his treatment of the problem fails to answer our concerns over the deep self in any substantive manner. In "Identification and Wholeheartedness," Frankfurt responds to Watson's criticism, drawing a parallel between decisiveness regarding desires and decisiveness regarding correctness in arithmetic. When solving arithmetic problems we check our correctness by doing the problem over again, perhaps in a different manner. This second attempt at the problem however, is no more guaranteed of correctness than is the first. Its order in the sequence of attempts imparts no special significance to it. We can continue this process ad infinitum, but Frankfurt notes that at some point we must simply "*decide for some reason*" to adopt a given result (Frankfurt, 1987, 36). Just what the reason is is unimportant; all that matters is that we endorse one answer as our own rather than throw up our hands and give up. Frankfurt's use of decision then is reminiscent of Wittgenstein's use of the term "practice" in *Philosophical Investigations*. Verification, be it of the correctness of a math problem or of the extent to which we identify with a first-order desire, comes to a halt at some point, at which time we simply *do*.

Provided we eschew the plausibility of a genuinely Cartesian foundation for this string of successive verifications, it would seem that

Frankfurt has succeeded in answering the challenge raised by infinite regress. To some extent this is correct; he has demonstrated that the threat of an infinite regress *occurring* is implausible. However, this does nothing to answer the concern behind why such a regress is *threatening* in the first place. The regress was threatening because it made clear the fact that each successive layer of desire could only control the desires below it if it itself was controlled by a another desire. Halting the regress, then, may satisfy our demands as philosophers for elegant theoretical structures, but it accomplishes no real work for the Frankfurt-Watson model of the deep self. The final level of desire is no more under one's control then is the first. By divorcing the deep self view of its infinite regress, Frankfurt also divorces it of its only hope of introducing moral responsibility into persons.

Flawed as it is, however, it may be that the deep self view is not so much categorically wrong as it is incomplete. In the article "Sanity and the Metaphysics of Responsibility", Susan Wolf attempts to rescue the deep self view by attaching a second criterion that must be met for an individual to gain moral responsibility. She points out that "not all the things necessary for...responsibility must be types of power and control. We may need simply to *be* a certain way, even through it is not within our power to determine whether we are that way or not." (Wolf, 55, italics hers). Wolf's insight is that for the deep self view to escape the infinite regress presented by levels of desire, there must be some fact distinct from our desires that grants persons some important degree of control over them. Moreover, since this fact imparts control, it must not be something that itself depends upon control by the self.

Wolf finds such a fact in the concept she refers to as "sanity." She defines "sanity" as the "minimally sufficient ability cognitively and normatively to recognize and appreciate the world for what it is" (Wolf, 56). Wolf is insisting here that moral responsibility requires persons to be in a certain relation to the world, such that they have access to certain facts about it. The facts in question are normative values. By possessing the ability to understand these values, a characteristic that Wolf notes may be equally related to having certain experiences as it is to fundamental aspects of persons, persons are able to "understand and appreciate right and wrong" (Wolf, 59). In short, this relationship allows persons access to morally relevant considerations that allow them to recognize the moral nature of their actions.

The advantage gained by requiring this connection to the world is that it would seem to ground individual's understanding of the world in a semi-objective (or at least inter-subjective) perspective. Focus is shifted from the self to the self's place in the world. Wolf argues that this perspectival shift is important because it alters the impact of deep will on the self. In the Frankfurt-Watson model of the deep self, the will is capable of examining one's basic desires and selecting those that it wished to hold as motivational factors. The self then can be seen as capable of self-revision. Unfortunately, the deep self view provides no reason to believe that these deeper values are ultimately unharried by the same factors influencing our more basic desires. As such, simple revision is insufficient grounds for responsibility. Wolf argues that what is necessary for such responsibility is that the self not simply be revised, but corrected. As Wittgenstein demonstrated in *Philosophical*

Investigations, correction demands an appeal to some standard beyond the self. This is the necessity of the sanity clause. Those persons Wolf would characterize as 'insane' lack the proper connection to the world such that they can recognize these features to which one must appeal in the correction process. Their revisions, however much in line they may be with the deep will or how different they may be from earlier revisions, cannot be said to move persons any closer to responsibility. Revision may re-create the self, but the recreated self holds no advantage over its original iteration.

Returning to the case of Jojo, one can see how the sanity condition moves the deep self view more in line with our intuitions. While JoJo's acts may be terribly wrong, the causal forces which shaped his person rendered him incapable, to no fault of his own, of seeing his actions as such. Likewise, the Wolfian sane deep self view goes a long way to explain similar moral intuitions, such as our reluctance to hold persons deeply responsible for actions in line with cultural values or those persons who genuinely are psychologically disordered.

Two points are important to note here. First, the sanity clause does not introduce a form of untenable cultural relativism. In no way does the sanity clause alter our judgements of the acts performed by persons such as JoJo. Widespread, culturally-endorsed wrongs such as racism or chauvinism are still immoral. Likewise, one is not guilty of an immoral act simply because it defies convention. While our expectations for persons vary according their histo-cultural situatedness, the moral status of their actions is an objective characteristic of the world. Secondly, it is worth reiterating that Wolf does not advance the sanity

clause as a means of confirming the worries with which this discussion began. While *some* persons may lack responsibility for their actions, for most of us, the deep selves we posses "unavoidably contain the ability to know right from wrong - we unavoidably do have the resources and reasons on which to base self-correction" (Wolf, 58).

But it's not entirely clear how Wolf might demonstrate that most of our deep selves do in fact posses the ability to know right from wrong. Her claim, after all, is not ethical but metaphysical; it is making an assertion about the relation between persons and the world. It would seem that the best defense of this assertion, then, is simply the fact that most persons ostensibly are moral. But if behavior is any measure of the relationship between person and world, which the sane deep self view would seem to demand, then we must consider the possibility that any immoral act amounts to evidence that the person in question lacks the type of connection requisite for moral responsibility.

Of course, any one piece of evidence cannot be considered conclusive. The performance of an immoral act alone is insufficient proof that an individual lacks the type of connection to the world required for moral responsibility. Acts must be considered in the context of the whole person, and in such light one aberrant behavior, however repulsive it may be, is not inherently representative of an individual failing to posses the correct relation to the world. For if persons are capable of recognizing the immoral nature of their wrong acts, they are in the correct relation to the world to be held responsible for it irregardless of the frequency or infrequency with which such acts are performed.

More important than the moral status of the act performed, then, would seem to be the agent's attitudes towards that action.

This is the point on which the sanity condition fails. For if persons' subjective attitudes towards their own acts determine their degree of sanity, then the very belief that one's actions are moral is sufficient to excuse one from responsibility for said acts. This is more deeply problematic for the sane deep self view than is immediately apparent. At first glance, the sanity clause would seem to largely fail the task to which Wolf set it. While it may succeed in rescuing some sense of moral responsibility, the domain of this type of responsibility is limited to those acts for which persons recognize their guilt before the act is committed. As such, those acts for which persons are morally responsible are relegated to what might broadly be conceived as mischief. Persons who knowingly commits a wrongful act because they simply desire it too much to resist are therefore morally responsible, while Hitler, Ted Kazynski, and members of the Klu Klux Klan are free from reproach, however appalling their acts might be.

Taken in isolation from other moral theory, these consequences of regarding sanity as a necessary condition for moral responsibility are not incoherent so much as highly counter-intuitive. This alone is insufficient grounds to dismiss sanity as a valuable consideration in determining moral responsibility. These consequences do demonstrate, however, that sanity fails when treated as a clause appended to the deep self view. The resulting sane deep self view is simply incoherent. As demonstrated above, the only persons who can be held morally responsible for wrongful acts are those who know that such acts are

wrong, but proceed with them none-the-less. Under the deep self view, such behavior implies that these persons possess the deep value that the act they are compelled to perform is wrong, but they cannot exert the force of will requisite to act on that value; in the end they are compelled by the shallow desire. But while the sanity condition demands that such persons are in fact responsible for their actions, the deep self view explicitly exempts them from responsibility. Cases such as this are no different from that of the unwilling addict who does not desire to take the drug, but cannot resist the compulsion to do just that. The deep self view therefore is not rescued by the sanity clause, but instead rendered deeply incoherent by it.

The sanity condition was doomed from the outset, however. The fundamental approach taken by the deep self view is flawed at its most basic level. However insightful Wolf's analysis is of the structural difficulties presented by the deep self view, the position cannot readily be patched. Wolf seems to have correctly located the source of the problem in the isolationist metaphysic of the deep self view. Frankfurt and Watson attempt to solve a problem about the relation between self and world by examining only the self. The intractability of this problem is made apparent through Wolf's sanity condition; for persons to acquire any deep culpability of the sort inherent to morality, a certain relationship is required between self and world. This relationship must be of a type such that the self alone cannot jerryrig the significant degree of control necessary for responsibility. In forcing the self to pull off such metaphysical slight of hand, Frankfurt and Watson fracture the self beyond recognition.

For what would be required of the deep self view to actually accomplish this task? Perhaps a more appropriate way to phrase this question, rather, is to ask how the Frankfurt-Watson model of the self could shelter the deep self from being compromised by the same causal factors that it readily grants a role in shaping our shallow desires? It would seem that at some level, the ability of the deep self to succeed here turns upon the fact that it must possess a different connection to the world than does the shallow self.[73] For there to be any differentiation in the causal predecessors of these two levels of self, this must minimally be the case. But this distinction in their relation to the world itself also places the different levels of self in a particular relation to each other. For if it is the case that they are distinct in their relations to the world they cannot be the same thing, but rather must be two separate things, each *external* to the other. Now there clearly is a sense in which this is intuitively correct. After all, we do at times claim to be "of two minds," and particularly in cases where conflicts arise between these two levels of desire (e.g., "I really don't want to eat that last slice of cake, but it just can't help myself.") there is a qualified sense in which we perceive ourselves as two selves in disagreement.

But when entertained as a serious model of *the* self, this fractured identity grates against our experiences of what it is like to be a person. Recall the structure of the self advanced by the deep self view. According to the Frankfurt-Watson model, it is our basic desires which actually motivate action, while the "real" self, the source of our

personhood, is buried in the deep levels of desire. This is plausible only if the different levels of self are connected in a deeply intimate fashion; if these basic desires really are part of our person. But if the deep self is truly external to these basic desires, which must be the case for it to accomplish the task set for it under the deep self view, then this cannot be the case. This model of persons describes action not as the direct, intimately-controlled affair that we experience, but as the proverbial donkey lead forever forward by a dangling carrot. Our person cannot directly invoke behaviors but instead must influence something external to itself and hope that it produces the desired results.

Of course, even if this deeply fractured model of the self was plausible, the deep self view would still not have succeeded in responding to the concerns with which the discussion began. For simply locating the deep self in a different relation to the world than the more shallow self does nothing but place the deep self under a different set of influences. Moreover, it is not enough to require that the relation between deep self and world be of a certain type; this is the approach taken by the sanity clause. Though such a requirement may be quite necessary for a successful model of the self, we have already demonstrated that the deep self view cannot make recourse to such a clause without rendering itself incoherent. Rather, the deep self view seems to demand that our deep selves be, in essence, Cartesian. This is not to suggest that the deep self view is committed to some form of dualism. The Frankfurt-Watson model does, however, seem committed

---

[73] My use of the term "self" here is not intended to suggest that the shallow self is necessarily its own person so much as to recognize the

to invoking some type of non-situated, objective subject at its core if it is to succeed in sheltering the deep self from the causal forces at the heart of our concern regarding responsibility. Such a self, of course, is highly implausible. Rather than respond to the concerns with which this discussion began, such a self would seem to deny them altogether. However untenable of a solution this may be, the deep self view lacks any clear alternative. It is, at its core, metaphysically flawed in a deep sense.

It would seem then that the Frankfurt-Watson deep self view is incapable of providing coherent grounds on which to base moral responsibility. Despite its ability to account for some aspects of will and responsibility, the deep self view ultimately fails to provide a complete, plausible account of these concepts due to the deeply fractured, and ultimately untenable metaphysics of self on which it is based. Moreover, contrary to arguments made by Susan Wolf, the deep self view cannot be made either ethically or metaphysically sound through the inclusion of a sanity clause. While the concept of sanity is itself is a worthwhile addition discussions of moral responsibility, it cannot be coherently appended to the deep self view as Wolf suggests. If we are to attain a compatibilist theory of moral responsibility, then, we must look beyond the deep self view.

---

fact that there must be something which perceives these shallow desires.

Bibliography

Frankfurt, Harry G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosphy, 68(1)*, 5-20.

Frankfurt, Harry G. (1987). Identification and Wholeheartedness. In F. Schoeman (Ed), *Responsibility, Character, and the Emotions (pp. 27-45).* New York: Cambridge.

Watson, Gary. (1975). Free Agency. *The Journal of Philosophy 72 (8),* 205-220.

Wolf, Susan. (1987). Sanity and the Metaphysics of Responsibility. In F. Schoeman (Ed), *Responsibility, Character, and the Emotions* (pp. 46-62). New York: Cambridge.