# Sanity and the Metaphysics of Responsibility

SUSAN WOLF

Philosophers who study the problems of free will and responsibility have an easier time than most in meeting challenges about the relevance of their work to ordinary, practical concerns. Indeed, philosophers who study these problems are rarely faced with such challenges at all, since questions concerning the conditions of responsibility come up so obviously and so frequently in everyday life. Under scrutiny, however, one might question whether the connections between philosophical and nonphilosophical concerns in this area are real.

In everyday contexts, when lawyers, judges, parents, and others are concerned with issues of responsibility, they know, or think they know, what in general the conditions of responsibility are. Their questions are questions of application: Does this or that particular person meet this or that particular condition? Is this person mature enough, or informed enough, or sane enough to be responsible? Was he or she acting under posthypnotic suggestion or under the influence of a mind-impairing drug? It is assumed, in these contexts, that normal, fully developed adult human beings are responsible beings. The questions have to do with whether a given individual falls within the normal range.

By contrast, philosophers tend to be uncertain about the general conditions of responsibility, and they care less about dividing the responsible from the nonresponsible agents than about determining whether, and if so why, any of us are ever responsible for anything at all.

In the classroom, we might argue that the philosophical concerns grow out of the nonphilosophical ones, that they take off where the nonphilosophical questions stop. In this way, we might convince our students that even if they are not plagued by the philosophical worries, they ought to be. If they worry about whether a person is mature enough, informed enough, and sane enough to be responsible, then they should worry about whether that person is metaphysically free enough, too.

The argument I make here, however, goes in the opposite direction. My aim is not to convince people who are interested in the apparently nonphilosophical conditions of responsibility that they should go on to worry about the philosophical conditions as well, but rather to urge those who already worry about the philosophical problems not to leave the more mundane, prephilosophical problems behind. In particular, I suggest that the mundane recognition that *sanity* is a condition of

Susan Wolf, "Sanity and the Metaphysics of Responsibility," pp. 46–62 from Ferdinand David Schoeman (ed.), *Responsibility, Character and the Emotions: New Essays in Moral Psychology.* New York: Cambridge University Press, 1988. Reprinted by permission of Cambridge University Press. Reprinted by permission of the author Susan Wolf.

responsibility has more to do with the murky and apparently metaphysical problems which surround the issue of responsibility than at first meets the eye. Once the significance of the condition of sanity is fully appreciated, at least some of the apparently insuperable metaphysical aspects of the problem of responsibility will dissolve.

My strategy is to examine a recent trend in philosophical discussions of responsibility, a trend that tries, but I think ultimately fails, to give an acceptable analysis of the conditions of responsibility. It fails due to what at first appear to be deep and irresolvable metaphysical problems. It is here that I suggest that the condition of sanity comes to the rescue. What at first appears to be an impossible requirement for responsibility – the requirement that the responsible agent have created her- or himself – turns out to be the vastly more mundane and non controversial requirement that the responsible agent must, in a fairly standard sense, be sane.

## Frankfurt, Watson, and Taylor

The trend I have in mind is exemplified by the writings of Harry Frankfurt, Gary Watson, and Charles Taylor. I will briefly discuss each of their separate proposals, and then offer a composite view that, while lacking the subtlety of any of the separate accounts, will highlight some important insights and some important blind spots they share.

In his seminal article "Freedom of the Will and the Concept of a Person,"[1] Harry Frankfurt notes a distinction between freedom of action and freedom of the will. A person has freedom of action, he points out, if she (or he) has the freedom to do whatever she wills to do – the freedom to walk or sit, to vote liberal or conservative, to publish a book or open a store, in accordance with her strongest desires. Even a person who has freedom of action may fail to be responsible for her actions, however, if the wants or desires she has the freedom to convert into action are themselves not subject to her control. Thus, the person who acts under post-hypnotic suggestion, the victim of brainwashing, and the kleptomaniac might all possess freedom of action. In the standard contexts in which these examples are raised, it is assumed that none of the individuals is locked up or bound. Rather, these individuals are understood to act on what, at one level at least, must be called *their own desires*. Their exemption from responsibility stems from the fact that their own desires (or at least the ones governing their actions) are not up to them. These cases may be described in Frankfurt's terms as cases of people who possess freedom of action, but who fail to be responsible agents because they lack freedom of the will.

Philosophical problems about the conditions of responsibility naturally focus on an analysis of this latter kind of freedom: What *is* freedom of the will, and under what conditions can we reasonably be thought to possess it? Frankfurt's proposal is to understand freedom of the will by analogy to freedom of action. As freedom of action is the freedom to do whatever one wills to do, freedom of the will is the freedom to will whatever one wants to will. To make this point clearer, Frankfurt introduces a distinction between first-order and second-order desires. First-order desires are desires to do or to have various things; second-order desires are desires about what desires to have or what desires to make effective in action. In order for an agent to have both freedom of action and freedom of the will, that agent must be capable of governing his or her actions by first-order desires *and* capable of governing his or her first-order desires by second-order desires.

Gary Watson's view of free agency[2] – free and responsible agency, that is – is similar to Frankfurt's in holding that an agent is responsible for an action only if the desires expressed by that action are of a particular kind. While Frankfurt identifies the right kind of desires as desires that are supported by second-order desires, however, Watson draws a distinction between "mere" desires, so to speak, and desires that are *values*. According to Watson, the difference between free action and unfree action cannot be analyzed by reference to the logical form of the desires from which these various actions arise, but rather must relate to a difference in the quality of their source. Whereas some of my desires are just appetites or conditioned responses I find myself "stuck with," others

are expressions of judgments on my part that the objects I desire are good. Insofar as my actions can be governed by the latter type of desire – governed, that is, by my values or valuational system – they are actions that I perform freely and for which I am responsible.

Frankfurt's and Watson's accounts may be understood as alternate developments of the intuition that in order to be responsible for one's actions, one must be responsible for the self that performs these actions. Charles Taylor, in an article entitled "Responsibility for Self,"[3] is concerned with the same intuition. Although Taylor does not describe his view in terms of different levels or types of desire, his view is related, for he claims that our freedom and responsibility depend on our ability to reflect on, criticize, and revise our selves. Like Frankfurt and Watson, Taylor seems to believe that if the characters from which our actions flowed were simply and permanently *given* to us, implanted by heredity, environment, or God, then we would be mere vehicles through which the causal forces of the world traveled, no more responsible than dumb animals or young children or machines. But like the others, he points out that, for most of us, our characters and desires are not so brutely implanted – or, at any rate, if they are, they are subject to revision by our own reflecting, valuing, or second-order desiring selves. We human beings – and as far as we know, only we human beings – have the ability to step back from ourselves and decide whether we are the selves we want to be. Because of this, these philosophers think, we are responsible for our selves and for the actions that we produce.

Although there are subtle and interesting differences among the accounts of Frankfurt, Watson, and Taylor, my concern is with features of their views that are common to them all. All share the idea that responsible agency involves something more than intentional agency. All agree that if we are responsible agents, it is not just because our actions are within the control of our wills, but because, in addition, our wills are not just psychological states *in* us, but expressions of characters that come *from* us, or that at any rate are acknowledged and affirmed *by* us. For Frankfurt, this means that our wills must be ruled by our second-order desires; for Watson, that our wills must be governable by our system of values: for Taylor, that our wills must issue from selves that are subject to self-assessment and redefinition in terms of a vocabulary of worth. In one way or another, all these philosophers seem to be saying that the key to responsibility lies in the fact that responsible agents are those for whom it is not just the case that their actions are within the control of their wills, but also the case that their wills are within the control of their *selves* in some deeper sense. Because, at one level, the differences among Frankfurt, Watson, and Taylor may be understood as differences in the analysis or interpretation of what it is for an action to be under the control of this deeper self, we may speak of their separate positions as variations of one basic view about responsibility: the *deep-self view.*

## The Deep-Self View

Much more must be said about the notion of a deep self before a fully satisfactory account of this view can be given. Providing a careful, detailed analysis of that notion poses an interesting, important, and difficult task in its own right. The degree of understanding achieved by abstraction from the views of Frankfurt, Watson, and Taylor, however, should be sufficient to allow us to recognize some important virtues as well as some important drawbacks of the deep-self view.

One virtue is that this view explains a good portion of our pretheoretical intuitions about responsibility. It explains why kleptomaniacs, victims of brainwashing, and people acting under posthypnotic suggestion may not be responsible for their actions, although most of us typically are. In the cases of people in these special categories, the connection between the agents' deep selves and their wills is dramatically severed – their wills are governed not by their deep selves, but by forces external to and independent from them. A different intuition is that we adult human beings can be responsible for our actions in a way that dumb animals, infants, and machines cannot. Here the explanation is not in terms of a split between these

beings' deep selves and their wills; rather, the point is that these beings *lack* deep selves altogether. Kleptomaniacs and victims of hypnosis exemplify individuals whose selves are *alienated* from their actions; lower animals and machines, on the other hand, do not have the sorts of selves from which actions *can* be alienated, and so they do not have the sort of selves from which, in the happier cases, actions can responsibly flow.

At a more theoretical level, the deep-self view has another virtue: It responds to at least one way in which the fear of determinism presents itself.

A naive reaction to the idea that everything we do is completely determined by a causal chain that extends backward beyond the times of our births involves thinking that in that case we would have no control over our behavior whatsoever. If everything is determined, it is thought, then what happens happens, whether we want it to or not. A common, and proper, response to this concern points out that determinism does not deny the causal efficacy an agent's desires might have on his or her behavior. On the contrary, determinism in its more plausible forms tends to affirm this connection, merely adding that as one's behavior is determined by one's desires, so one's desires are determined by something else.[4]

Those who were initially worried that determinism implied fatalism, however, are apt to find their fears merely transformed rather than erased. If our desires are governed by something else, they might say, they are not *really* ours after all – or, at any rate, they are ours in only a superficial sense.

The deep-self view offers an answer to this transformed fear of determinism, for it allows us to distinguish cases in which desires are determined by forces foreign to oneself from desires which are determined *by* one's self – by one's "real," or second-order desiring, or valuing, or deep self, that is. Admittedly, there are cases, like that of the kleptomaniac or the victim of hypnosis, in which the agent acts on desires that "belong to" him or her in only a superficial sense. But the proponent of the deep-self view will point out that even if determinism is true, ordinary adult human action can be distinguished from this. Determinism implies that the desires which govern our actions are in turn governed by something else, but that something else will, in the fortunate cases, be our own deeper selves.

This account of responsibility thus offers a response to our fear of determinism; but it is a response with which many will remain unsatisfied. Even if my actions are governed by my desires and my desires are governed by my own deeper self, there remains the question: Who, or what, is responsible for this deeper self? The response above seems only to have pushed the problem further back.

Admittedly, some versions of the deep-self view, including Frankfurt's and Taylor's seem to anticipate this question by providing a place for the ideal that an agent's deep self may be governed by a still deeper self. Thus, for Frankfurt, second-order desires may themselves be governed by third-order desires, third-order desires by fourth-order desires, and so on. Also, Taylor points out that, as we can reflect on and evaluate our prereflective selves, so we can reflect on and evaluate the selves who are doing the first reflecting and evaluating, and so on. However, this capacity to recursively create endless levels of depth ultimately misses the criticism's point.

First of all, even if there is no *logical* limit to the number of levels of reflection or depth a person may have, there is certainly a psychological limit – it is virtually impossible imaginatively to conceive a fourth-, much less an eighth-order, desire. More important, no matter how many levels of self we posit, there will still, in any individual case, be a last level – a deepest self about whom the question "What governs it?" will arise, as problematic as ever. If determinism is true, it implies that even if my actions are governed by my desires, and my desires are governed by my deepest self, my deepest self will still be governed by something that must, logically, be external to myself altogether. Though I can step back from the values my parents and teachers have given me and ask whether these are the values I really want, the "I" that steps back will itself be a product of the parents and teachers I am questioning.

The problem seems even worse when one sees that one fares no better if determinism is

false. For if my deepest self is not determined by something external to myself, it will still not be determined by *me*. Whether I am a product of carefully controlled forces or a result of random mutations, whether there is a complete explanation of my origin or no explanation at all, *I* am not, in any case, responsible for my existence; I am not in control of my deepest self.

Thus, though the claim that an agent is responsible for only those actions that are within the control of his or her deep self correctly identifies a necessary condition for responsibility – a condition that separates the hypnotized and the brainwashed, the immature and the lower animals from ourselves, for example – it fails to provide a sufficient condition of responsibility that puts all fears of determinism to rest. For one of the fears invoked by the thought of determinism seems to be connected to its implication that we are but intermediate links in a causal chain, rather than ultimate, self-initiating sources of movement and change. From the point of view of one who has this fear, the deep-self view seems merely to add loops to the chain, complicating the picture but not really improving it. From the point of view of one who has this fear, responsibility seems to require being a prime mover unmoved, whose deepest self is itself neither random *nor* externally determined, but is rather determined *by* itself – who is, in other words, self-created.

At this point, however, proponents of the deep-self view may wonder whether this fear is legitimate. For although people evidently can be brought to the point where they feel that responsible agency requires them to be ultimate sources of power, to the point where it seems that nothing short of self-creation will do, a return to the internal standpoint of the agent whose responsibility is in question makes it hard to see what good this metaphysical status is supposed to provide or what evil its absence is supposed to impose.

From the external standpoint, which discussions of determinism and indeterminism encourage us to take up, it may appear that a special metaphysical status is required to distinguish us significantly from other members of the natural world. But proponents of the deep-self view will suggest this is an illusion that a return to the internal standpoint should dispel. The possession of a deep self that is effective in governing one's actions is a sufficient distinction, they will say. For while other members of the natural world are not in control of the selves that they are, we, possessors of effective deep selves, are in control. We can reflect on what sorts of beings we are, and on what sorts of marks we make on the world. We can change what we don't like about ourselves, and keep what we do. Admittedly, we do not create ourselves from nothing. But as long as we can revise ourselves, they will suggest, it is hard to find reason to complain. Harry Frankfurt writes that a person who is free to do what he wants to do and also free to want what he wants to want has "all the freedom it is possible to desire or to conceive."[5] This suggests a rhetorical question: If you are free to control your actions by your desires, and free to control your desires by your deeper desires, and free to control those desires by still deeper desires, what further kind of freedom can you want?

## The Condition of Sanity

Unfortunately, there is a further kind of freedom we can want, which it is reasonable to think necessary for responsible agency. The deep-self view fails to be convincing when it is offered as a complete account of the conditions of responsibility. To see why, it will be helpful to consider another example of an agent whose responsibility is in question.

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.

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.

In light of JoJo's heritage and upbringing – both of which he was powerless to control – it is dubious at best that he should be regarded as responsible for what he does. 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. However, note that JoJo is someone whose actions are controlled by his desires and whose desires are the desires he wants to have: That is, his actions are governed by desires that are governed by and expressive of his deepest self.

The Frankfurt – Watson – Taylor strategy that allowed us to differentiate our normal selves from the victims of hypnosis and brainwashing will not allow us to differentiate ourselves from the son of Jo the First. In the case of these earlier victims, we were able to say that although the actions of these individuals were, at one level, in control of the individuals themselves, these individuals themselves, qua agents, were not the selves they more deeply wanted to be. In this respect, these people were unlike our happily more integrated selves. However, we cannot say of JoJo that his self, qua agent, is not the self he wants it to be. It *is* the self he wants it to be. From the inside, he feels as integrated, free, and responsible as we do.

Our judgment that JoJo is not a responsible agent is one that we can make only from the outside – from reflecting on the fact, it seems, that his deepest self is not up to him. Looked at from the outside, however, our situation seems no different from his – for in the last analysis, it is not up to any of us to have the deepest selves we do. Once more, the problem seems metaphysical – and not just metaphysical, but insuperable. For, as I mentioned before, the problem is independent of the truth of determinism. Whether we are determined or undetermined, we cannot have created our deepest selves. Literal self-creation is not just empirically, but logically impossible.

If JoJo is not responsible because his deepest self is not up to him, then we are not responsible either. Indeed, in that case responsibility would be impossible for anyone to achieve. But I believe the appearance that literal self-creation is required for freedom and responsibility is itself mistaken.

The deep-self view was right in pointing out that freedom and responsibility require us to have certain distinctive types of control over our behavior and our selves. Specifically, our actions need to be under the control of our selves, and our (superficial) selves need to be under the control of our deep selves. Having seen that these types of control are not enough to guarantee us the status of responsible agents, we are tempted to go on to suppose that we must have yet another kind of control to assure us that even our deepest selves are somehow up to us. But not all the things necessary for freedom and responsibility must be types of power and control. We may need simply to *be* a certain way, even though it is not within our power to determine whether we are that way or not.

Indeed, it becomes obvious that at least one condition of responsibility is of this form as soon as we remember what, in everyday contexts, we have known all along – namely, that in order to be responsible, an agent must be *sane*. It is not ordinarily in our power to determine whether we are or are not sane. Most of us, it would seem, are lucky, but some of us are not. Moreover, being sane does not necessarily mean that one has any type of power or control an insane person lacks. Some insane people, like JoJo and some actual political leaders who resemble him, may have complete control of their actions, and even complete control of their acting selves. The desire to be sane is thus not a desire for another form of control; it is rather a desire that one's self be connected to the world in a certain way – we could even say it is a desire that one's self be *controlled by* the world in certain ways and not in others.

This becomes clear if we attend to the criteria for sanity that have historically been dominant in legal questions about responsibility. According to the M'Naughten Rule, 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. Insofar as one's desire to be sane involves a desire to know what one is

doing – or more generally, a desire to live in the real world – it is a desire to be a controlled (to have, in this case, one's *beliefs* controlled) by perceptions and sound reasoning that produce an accurate conception of the world, rather than by blind or distorted forms of response. The same goes for the second constituent of sanity – only, in this case, one's hope is that one's *values* be controlled by processes that afford an accurate conception of the world.[6] Putting these two conditions together, we may understand sanity, then, as the minimally sufficient ability cognitively and normatively to recognize and appreciate the world for what it is.

There are problems with this definition of sanity, at least some of which will become obvious in what follows, that make it ultimately unacceptable either as a gloss on or an improvement of the meaning of the term in many of the contexts in which it is used. The definition offered does seem to bring out the interest sanity has for us in connection with issues of responsibility, however, and some pedagogical as well as stylistic purposes will be served if we use sanity hereafter in this admittedly specialized sense.

## The Sane Deep-Self View

So far I have argued that the conditions of responsible agency offered by the deep-self view are necessary but not sufficient. Moreover, the gap left open by the deep-self view seems to be one that can be filled only by a metaphysical, and, as it happens, metaphysically impossible addition. I now wish to argue, however, that the condition of sanity, as characterized above, is sufficient to fill the gap. In other words, the deep-self view, supplemented by the condition of sanity, provides a satisfying conception of responsibility. The conception of responsibility I am proposing, then, agrees with the deep-self view in requiring that a responsible agent be able to govern her (or his) actions by her desires and to govern her desires by her deep self. In addition, my conception insists that the agent's deep self be sane, and claims that this is *all* that is needed for responsible agency. By contrast to the plain deep-self view, let us call this new proposal the *sane deep-self view.*

It is worth noting, to begin with, that this new proposal deals with the case of JoJo and related cases of deprived childhood victims in ways that better match our pretheoretical intuitions. Unlike the plain deep-self view, the sane deep-self view offers a way of explaining why JoJo is not responsible for his actions without throwing our own responsibility into doubt. For, although like us, JoJo's actions flow from desires that flow from his deep self, unlike us, JoJo's deep self is itself insane. Sanity, remember, involves the ability to know the difference between right and wrong, and a person who, even on reflection, cannot see that having someone tortured because he failed to salute you is wrong plainly lacks the requisite ability.

Less obviously, but quite analogously, this new proposal explains why we give less than full responsibility to persons who, though acting badly, act in ways that are strongly encouraged by their societies – the slaveowners of the 1850s, the Nazis of the 1930s, and many male chauvinists of our fathers' generation, for example. These are people, we imagine, who falsely believe that the ways in which they are acting are morally acceptable, and so, we may assume, their behavior is expressive of or at least in accordance with these agents' deep selves. But their false beliefs in the moral permissibility of their actions and the false values from which these beliefs derived may have been inevitable, given the social circumstances in which they developed. If we think that the agents could not help but be mistaken about their values, we do not blame them for the actions those values inspired?[7]

It would unduly distort ordinary linguistic practice to call the slaveowner, the Nazi, or the male chauvinist even partially or locally insane. Nonetheless, the reason for withholding blame from them is at bottom the same as the reason for withholding it from JoJo. Like JoJo, they are, at the deepest level, unable cognitively and normatively to recognize and appreciate the world for what it is. In our sense of the term, their deepest selves are not fully *sane.*

The sane deep-self view thus offers an account of why victims of deprived childhoods as well as victims of misguided societies may not be responsible for their actions, without implying that we are not responsible for ours.

The actions of these others are governed by mistaken conceptions of value that the agents in question cannot help but have. Since, as far as we know, our values are not, like theirs, unavoidably mistaken, the fact that these others are not responsible for their actions need not force us to conclude that we are not responsible for ours.

But it may not yet be clear why sanity, in this special sense, should make such a difference – why, in particular, the question of whether someone's values are unavoidably *mistaken* should have any bearing on their status as responsible agents. The fact that the sane deep-self view implies judgments that match our intuitions about the difference in status between characters like JoJo and ourselves provides little support for it if it cannot also defend these intuitions. So we must consider an objection that comes from the point of view we considered earlier which rejects the intuition that a relevant difference can be found.

Earlier, it seemed that the reason JoJo was not responsible for his actions was that although his actions were governed by his deep self, his deep self was not up to him. But this had nothing to do with his deep self's being mistaken or not mistaken, evil or good, insane or sane. If JoJo's values are unavoidably mistaken, our values, even if not mistaken, appear to be just as unavoidable. When it comes to freedom and responsibility, isn't it the unavoidability, rather than the mistakenness, that matters?

Before answering this question, it is useful to point out a way in which it is ambiguous: The concepts of avoidability and mistakenness are not unequivocally distinct. One may, to be sure, construe the notion of avoidability in a purely meta-physical way. Whether an event or state of affairs is unavoidable under this construal depends, as it were, on the tightness of the causal connections that bear on the event's or state of affairs' coming about. In this sense, our deep selves do seem as unavoidable for us as JoJo's and the others' are for them. For presumably we are just as influenced by our parents, our cultures, and our schooling as they are influenced by theirs. In another sense, however, our characters are not similarly unavoidable.

In particular, in the cases of JoJo and the others, there are certain features of their characters that they cannot avoid *even though these features are seriously mistaken, misguided, or bad.* This is so because, in our special sense of the term, these characters are less than fully sane. Since these characters lack the ability to know right from wrong, they are unable to revise their characters on the basis of right and wrong, and so their deep selves lack the resources and the reasons that might have served as a basis for self-correction. Since the deep selves *we* unavoidably have, however, are sane deep selves – deep selves, that is, that unavoidably *contain* the ability to know right from wrong – we unavoidably do have the resources and reasons on which to base self-correction. What this means is that though in one sense we are no more in control of our deepest selves than JoJo et al., it does not follow in our case, as it does in theirs, that we would be the way we are, even if it is a bad or wrong way to be. However, if this does not follow, it seems to me, our absence of control at the deepest level should not upset us.

Consider what the absence of control at the deepest level amounts to for us: Whereas JoJo is unable to control the fact that, at the deepest level, he is not fully sane, we are not responsible for the fact that, at the deepest level, we are. It is not up to us to *have* minimally sufficient abilities cognitively and normatively to recognize and appreciate the world for what it is. Also, presumably, it is not up to us to have lots of other properties, at least to begin with – a fondness for purple, perhaps, or an antipathy for beets. As the proponents of the plain deep-self view have been at pains to point out, however, we do, if we are lucky, have the ability to revise our selves in terms of the values that are held by or constitutive of our deep selves. If we are lucky enough both to have this ability and to have our deep selves be sane, it follows that although there is much in our characters that we did not choose to have, there is nothing irrational or objectionable in our characters that we are compelled to keep.

Being sane, we are able to understand and evaluate our characters in a reasonable way, to notice what there is reason to hold on to, what there is reason to eliminate, and what, from a

rational and reasonable standpoint, we may retain or get rid of as we please. Being able as well to govern our superficial selves by our deep selves, then, we are able to change the things we find there is reason to change. This being so, it seems that although we may not be *metaphysically* responsible for ourselves – for, after all, we did not create ourselves from nothing – we are *morally* responsible for ourselves, for we are able to understand and appreciate right and wrong, and to change our characters and our actions accordingly.

## Self-Creation, Self-Revision, and Self-Correction

At the beginning of this chapter, I claimed that recalling that sanity was a condition of responsibility would dissolve at least some of the appearance that responsibility was metaphysically impossible. To see how this is so, and to get a fuller sense of the sane deep-self view, it may be helpful to put that view into perspective by comparing it to the other views we have discussed along the way.

As Frankfurt, Watson, and Taylor showed us, in order to be free and responsible we need not only to be able to control our actions in accordance with our desires, we need to be able to control our desires in accordance with our deepest selves. We need, in other words, to be able to *revise* ourselves – to get rid of some desires and traits, and perhaps replace them with others on the basis of our deeper desires or values or reflections. However, consideration of the fact that the selves who are doing the revising might themselves be either brute products of external forces or arbitrary outputs of random generation made us wonder whether the capacity for self-revision was enough to assure us of responsibility – and the example of JoJo added force to the suspicion that it was not. Still, if the ability to revise ourselves is not enough, the ability to create ourselves does not seem necessary either. Indeed, when you think of it, it is unclear why anyone should want self-creation. Why should anyone be disappointed at having to accept the idea that one has to get one's start somewhere? It is an idea that most of us have lived with quite contentedly all along. What we do have reason to want, then, is something more than the ability to revise ourselves, but less than the ability to create ourselves. Implicit in the sane deep-self view is the idea that what is needed is the ability to *correct* (or improve) ourselves.

Recognizing that in order to be responsible for our actions, we have to be responsible for our selves, the sane deep-self view analyzes what is necessary in order to be responsible for our selves as (1) the ability to evaluate ourselves sensibly and accurately, and (2) the ability to transform ourselves insofar as our evaluation tells us to do so. We may understand the exercise of these abilities as a process whereby we *take* responsibility for the selves that we are but did not ultimately create. The condition of sanity is intrinsically connected to the first ability; the condition that we able to control our superficial selves by our deep selves is intrinsically connected to the second.

The difference between the plain deep-self view and the sane deep-self view, then, is the difference between the requirement of the capacity for self-revision and the requirement of the capacity for self-correction. Anyone with the first capacity can *try* to take responsibility for himself or herself. However, only someone with a sane deep self – a deep self that can see and appreciate the world for what it is – can self-evaluate sensibly and accurately. Therefore, although insane selves can try to take responsibility for themselves, only sane selves will properly be accorded responsibility.

## Two Objections Considered

At least two problems with the sane deep-self view are so glaring as to have certainly struck many readers. In closing, I shall briefly address them. First, some will be wondering how, in light of my specialized use of the term "sanity," I can be so sure that "we" are any saner than the nonresponsible individuals I have discussed. What justifies my confidence that, unlike the slaveowners, Nazis, and male chauvinists, not to mention JoJo himself, we are able to understand and appreciate the world for what it is? The answer to this is that nothing justifies this except widespread intersubjective

agreement and the considerable success we have in getting around in the world and satisfying our needs. These are not sufficient grounds for the smug assumption that we are in a position to see the truth about *all* aspects of ethical and social life. Indeed, it seems more reasonable to expect that time will reveal blind spots in our cognitive and normative outlook, just as it has revealed errors in the outlooks of those who have lived before. But our judgments of responsibility can only be made from here, on the basis of the understandings and values that we can develop by exercising the abilities we do possess as well and as fully as possible.

If some have been worried that my view implicitly expresses an overconfidence in the assumption that we are sane and therefore right about the world, others will be worried that my view too closely connects sanity with being right about the world, and fear that my view implies that anyone who acts wrongly or has false beliefs about the world is therefore insane and so not responsible for his or her actions. This seems to me to be a more serious worry, which I am sure I cannot answer to everyone's satisfaction.

First, it must be admitted that the sane deep-self view embraces a conception of sanity that is explicitly normative. But this seems to me a strength of that view, rather than a defect. Sanity *is* a normative concept, in its ordinary as well as in its specialized sense, and severely deviant behavior, such as that of a serial murderer of a sadistic dictator, does constitute evidence of a psychological defect in the agent. The suggestion that the most horrendous, stomach-turning crimes could be committed only by an insane person – an inverse of Catch-22, as it were – must be regarded as a serious possibility, despite the practical problems that would accompany general acceptance of that conclusion.

But, it will be objected, there is no justification, in the sane deep-self view, for regarding only horrendous and stomach-turning crimes as evidence of insanity in its specialized sense. If sanity is the ability cognitively and normatively to understand and appreciate the world for what it is, then *any* wrong action or false belief will count as evidence of the absence of that ability. This point may also be granted, but we must be careful about what conclusion to draw. To be sure, when someone acts in a way that is not in accordance with acceptable standards of rationality and reasonableness, it is always appropriate to look for an explanation of why he or she acted that way. The hypothesis that the person was unable to understand and appreciate that an action fell outside acceptable bound will always be a possible explanation. Bad performance on a math test always suggests the possibility that the testee is stupid. Typically, however, other explanations will be possible, too – for example, that the agent was too lazy to consider whether his or her action was acceptable, or too greedy to care, or, in the case of the math testee, that he or she was too occupied with other interests to attend class or study. Other facts about the agent's history will help us decide among these hypotheses.

This brings out the need to emphasize that sanity, in the specialized sense, is defined as the *ability* cognitively and normatively to understand and appreciate the world for what it is. According to our commonsense understandings, having this ability is one thing and exercising it is another – at least some wrong-acting, responsible agents presumably fall within the gap. The notion of "ability" is notoriously problematic, however, and there is a long history of controversy about whether the truth of determinism would show our ordinary ways of thinking to be simply confused on this matter. At this point, then, metaphysical concerns may voice themselves again – but at least they will have been pushed into a narrower, and perhaps a more manageable, corner.

The sane deep-self view does not, then, solve all the philosophical problems connected to the topics of free will and responsibility. If anything, it highlights some of the practical and empirical problems, rather than solves them. It may, however, resolve some of the philosophical, and particularly, some of the metaphysical problems, and reveal how intimate are the connections between the remaining philosophical problems and the practical ones.

1. Harry Frankfurt, "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* LXVIII (1971), 5–20.
2. Gary Watson, "Free Agency," *Journal of Philosophy* LXXII (1975), 205–20.
3. Charles Taylor, "Responsibility for Self," in A. E. Rorty, ed., *The Identities of Persons* (Berkeley: University of California Press, 1976), pp. 381–99.
4. See, e.g., David Hume. *A Treatise of Human Nature* (Oxford: Oxford University Press, 1967), pp. 399–406, and R. E. Hobart, "Free Will as Involving Determination and Inconceivable Without It," *Mind* 43 (1934).
5. Frankfurt, p. 16.
6. Strictly speaking, perception and sound reasoning may not be enough to ensure the ability to achieve an accurate conception of what one is doing and especially to achieve a reasonable normative assessment of one's situation. Sensitivity and exposure to certain realms of experience may also be necessary for these goals. For the purpose of this essay, I understand "sanity" to include whatever it takes to enable one to develop an adequate conception of one's world. In other contexts, however, this would be an implausibly broad construction of the term.
7. Admittedly, it is open to question whether these individuals were in fact unable to help having mistaken values, and indeed, whether recognizing the errors of their society would even have required exceptional independence or strength of mind. This is presumably an empirical question, the answer to which is extraordinarily hard to determine. My point here is simply that *if* we believe they are unable to recognize that their values are mistaken, we do not hold them responsible for the actions that flow from these values, and *if* we believe their ability to recognize their normative errors is impaired, we hold them less than fully responsible for relevant actions.