# Moral Sanity Reformulated: Revising Susan Wolf's Sanity-Condition

*Benjamin Edelson*

## I. Introduction

Some people think that moral responsibility is a metaphysical impossibility because the universe is causally determined. Others think that determinism must be false because we know *a priori* that free will (and therefore responsibility) exists. A third group sets itself apart from the first two in its rejection of determinism's relevance to the issue of moral responsibility at all. On this view, responsibility is made possible by certain psychological capacities, capacities which either exist or do not irrespective of the truth or falsity of determinism. So conceived, moral responsibility is *compatible* with a deterministic universe. The question of what exactly the pertinent capacities are, however, is the subject of ongoing debate among compatibilists. Several influential answers involve what Susan Wolf labels the 'deep-self view' – the idea one's will must be connected to some deep or ultimate manifestation of one's *self*. In her paper 'Sanity and the metaphysics of responsibility,' Wolf takes issue with the deep-self view, suggesting that there is a further condition to be met: *sanity*. Sane agents have the capacity to "cognitively and normatively recognize and appreciate the world for what it is."[42] Just as their empirical beliefs must reflect the world's physical reality, Wolf thinks, so must their values accurately reflect its moral reality.

Building on Wolf's critique of the deep-self compatibilists, I will offer what I think is a necessary revision to her so-called 'sane deep-self view.' While there is promise in looking to sanity as the necessary capacity for moral responsibility, I think Wolf errs in emphasizing the *substance* of one's moral values as a benchmark for sanity. This misplaced focus prevents Wolf's theory from being able to account for changes in genuinely thought-through values over time, as well as differences between values sincerely held by contemporary agents. As a result, it fails to accurately encapsulate our real-life responsibility-practices. A better articulation of the sane deep-self view would

[42] Wolf, Susan (1987). 'Sanity and the metaphysics of responsibility.' In *Responsibility, Character, and the Emotions: New Essays in Moral Psychology.* Edited by Ferdinand Schoeman. Cambridge University Press, pg. 56.

focus more on the capacity to justify one's actions by citing *any* general behavioral principle, and less on the particulars of the principles themselves.

Wolf "embraces a conception of sanity that is explicitly normative."[43] This, I will argue, is her problem. My task is to conceive of moral sanity in a way that is *not* normative. In so doing, I hope to patch some of the sane deep-self view's holes and make it a more plausible candidate in the compatibilists' search for the capacity necessary for responsibility. I will first present Wolf's formulation of the sanity-condition. Then I will point out its problematic implications, suggest and defend my fix, and address potential issues with my proposal.

## II. Wolf's Sanity-Condition

Wolf arrives at her conception of sanity via her dissatisfaction with Harry Frankfurt's view of responsibility. Frankfurt thinks that the distinguishing mark between agents and non-agents is the former's capacity not just to do as they want, but to critically reflect on those wants *and* structure their will accordingly. The capacity for 'second-order desires' – a desire "simply to have a certain desire"[44] – is not enough, for, as the author shows, there are agents who meet this criterion whom we would regard as poor candidates for responsibility. He offers the example of a 'willing' drug addict – someone who struggles against his addiction, but is not capable of caring "whether his craving or aversion gets the upper hand."[45] This addict, being "neutral with regard to the conflict between his desire to take the drug and his desire to refrain from taking it,"[46] lacks a capacity key to responsibility: the ability to "[want] a certain desire to be his will,"[47] or the freedom to

[43] Ibid, pg. 61.

[44] Frankfurt, Harry (1971). 'Freedom of the Will and the Concept of a Person.' In *The Journal of Philosophy*, Vol. 68, No. 1, pg. 10.

[45] Ibid, pg. 13.

[46] Ibid, pg. 12.

[47] Ibid, pg. 10.

"want what he wants to want."[48] Frankfurt labels this higher-level connection between one's desires and their will 'second-order volitions.' Responsible agents' "wills are within the control of their *selves* in some deeper sense"[49] – they 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."[50]

But the question remains: "Who, or what, is responsible for this deeper self?"[51] Why stop at second-order volitions? Why are not third-, fourth-, or fifth-order volitions necessary for responsibility? We are seemingly no more responsible for our second-order volitions than we are for our first-order ones.

Wolf answers by suggesting that, to really have second-order volitions, we must be able to direct our will in pursuit of the *correct* kinds of ends. For agents to "understand and evaluate their 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,"[52] they must possess "the ability cognitively and normatively to understand and appreciate the world for what it is."[53] *Cognitively* in that they can recognize a chair for a chair, and *normatively* in that they can recognize right from wrong. If agents are to correct their desires and wills in accordance with the world's normative makeup, their normative beliefs about the world must be correct – they must be *sanely* connected to the world. So, to be properly held responsible one's deep self need be sane.

[48] Ibid, pg. 15.

[49] Wolf, pg. 50.

[50] Ibid, pg. 49.

[51] Ibid, pg. 51.

[52] Ibid, pg. 59.

[53] Ibid, pg. 62.

Wolf thinks that this “explains why we give less than full responsibility to persons who, though acting badly, act in ways that are strongly encouraged by their societies...many male chauvinists of our fathers’ generation, for example.”[54] She acknowledges that it “would unduly distort ordinary linguistic practice to call...the male chauvinist even partially or locally insane,” but, despite this, maintains that they indeed are insane in that the normative basis for their sexism is so terribly mistaken that it demonstrates a lack of capacity to grasp the objective moral makeup of the world. In that sense they are insane; they simply cannot appreciate reality.

Here a glaring question arises: “What justifies [Wolf’s] confidence that, unlike the slaveholders, Nazis and male chauvinists…we are able to understand and appreciate the world for what it is?”[55] The debate between those who think ethical truths are objective and those who think they are subjective has a long history. But Wolf wisely avoids wading into that disagreement in any substantive way. Instead, she simply asserts that “nothing justifies this [confidence] except wide intersubjective agreement and the considerable success we have in getting around the world and satisfying our needs.”[56] We will undoubtedly continue to revise and improve on our values going forwards. But it seems to her that we have a fundamental normative understanding of the world that Nazis and chauvinists lack.

## III. Wolf’s Problems

The first issue with the sanity-condition is its implication that, whenever wide intersubjective agreement about proper norms of behavior shifts (as it has over time, and no doubt will continue to), the conditions for sanity also shift. Wolf is confident that we are sane today, but by her criteria we could legitimately be called insane by the people of

[54] Ibid, pg. 57.

[55] Ibid, pg. 60.

[56] Ibid.

tomorrow. For example: according to some wide intersubjective moral agreement of the 1950s, contemporary chauvinists could be said to be genuinely self-correcting when they examined and reaffirmed their sexist values. Now, we find the chauvinists' introspection processes objectionable. We think they came to the wrong conclusion, and have our own thought-through reasons for believing this. In another 100 years, if wide intersubjective moral agreement shifts in favor of sexism, people might think the chauvinists were correct in their defense of their values. Such a shift, while perhaps improbable, is entirely plausible.

But the above means that in the 50s chauvinists *were* sane, are currently *insane*, and in the future they *will* be sane again. How could this be so if sanity is just the ability to recognize the world for what it objectively *is*? Surely the world's objective makeup has not changed since the 50s.

To ask this is not necessarily to argue against moral objectivity. It is merely to point out an inability to reconcile Wolf's standard of sanity – values endorsed by wide intersubjective agreement – with radical shifts in such agreement over time. For example: many people currently deeply disagree on the morality of euthanasia. Both camps have rigorous moral arguments for their respective positions. If, in 100 years, euthanasia is widely recognized as seriously unethical, then, on Wolf's account, the people of the future would be justified in regarding today's euthanasia-defenders as "unable [to] normatively recognize and appreciate the world for what it is" and therefore "not fully *sane*."[57]

But of course many euthanasia-defenders *are* sane. They are sane because they are capable of justifying their view by engaging in good-faith deliberation about how people should behave. Their sanity is not a function of which side of the euthanasia

[57] Ibid, pg. 57.

debate they fall on. Wolf makes the particular values one holds determinative of sanity, but is unable to provide any substantive test for which values are the 'sane' ones.

Consensuses also vary (radically) in different locations and cultures around the globe. Does *wide* refer to a given community, country, continent – or the entire world? Even if we could define the area, would we need 51%, 69%, or 82% agreement for a particular moral view to be 'objectively' sane? On the very contentious issues there is never 100% concurrence. And even on the less controversial ones there usually exist many different consensuses at a given time.

The second problem with Wolf's condition is that it eliminates the viability of genuine moral disagreement, which is a fundamental part of moral thought. On her view, whom may we validly hold responsible? Only, it seems, people who share our (objectively correct) values, but fail to live up to them. But this rather limited category does not include many types of agents we actually want to hold responsible. Wolf addresses this towards the end of her paper, admitting that her 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."[58] For, "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."[59] She answers by suggesting that "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."[60] Perhaps the agent has the capacity to recognize the objectively correct values, and so is sane, but simply fell short of acting upon those values because of other factors. In many cases, this response will suffice. But it will not help when we want

[58] Ibid, pg. 61.

[59] Ibid.

[60] Ibid.

to assign responsibility in cases where we *genuinely* morally disagree – in cases where both sides have indeed thought their values and positions through, and are committed to defending them. In fact, these cases are often the ones in which we are most desperate to morally blame.

The real trouble for Wolf's view arises in cases in which neither side is being sloppy, yet both are genuinely convinced that they are understanding and appreciating the world's normative makeup for what it is. "The suggestion that the most horrendous, stomach-turning crimes could only be committed by an insane person," Wolf writes, "must be regarded as a serious possibility, despite the practical problems that would accompany general acceptance of that conclusion."[61] The issue is precisely that in certain situations there is serious disagreement about what constitutes such crimes. To many anti-speciesists there is a 'Holocaust on Your Plate' every time you dig into a meal of steak (think 'MEAT IS MURDER!').[62] And yet there are other long-standing philosophical arguments explaining why eating non-human animals is morally permissible. Wolf's view implies that one camp is objectively morally insane. But anyone who has talked to thoughtful representatives from both these camps knows that is untrue. Ethical deliberation is difficult, and clearly-thinking people arrive at divergent conclusions. But this does not make them insane. If it did, we would have no way of knowing on which issues we currently hold sane or insane views – and yet Wolf insists that we are sane in most of our views.

To return to euthanasia: the opposing positions are marked by affirmations of two different moral judgments. Euthanasia-attackers endorse *A*: 'Life is intrinsically good, so one ought not kill.' And euthanasia-defenders endorse *B*: 'Life is good insofar as people enjoy it, so one ought not kill those who want to go on living.'

[61] Ibid.

[62] Hamilton, Jill (n.d.). 'Ethics Case Studies: Using the 'Holocaust' Metaphor.' *Society of Professional Journalists.*

Per Wolf, each camp should regard their respective opponents with a puzzling sort of moral indifference. ‘We may genuinely disagree,’ the attackers would be expected to say, ‘but all that means is that in endorsing *B* you demonstrate an inability to grasp the objective normative makeup of the world. You are morally insane; therefore, it is unfair for me to hold you morally responsible for your actions when you enable people to commit euthanasia, even though they are knowingly committing ‘*horrendous, stomach-turning crimes*.’’

No one would address their normative opponent like this. The euthanasia-attacker would actually say: ‘We genuinely disagree, and your endorsement of *B* is mistaken for *x* reasons. You are morally wrong; therefore, I will hold you morally responsible for enabling people to end their lives.’ For the attacker, the defender is a prime candidate for moral blame, precisely *because* they have the ‘wrong’ values.

That is why we want to hold Nazis, chauvinists, and slaveholders responsible. It is not because they hold the ‘right’ values, but fail to put them into practice – it is because they thinkingly endorse the ‘wrong’ values. This confusion is the reason for Wolf’s distortion of “ordinary linguistic practice.”[63] She correctly notes that philosophical reflection about words’ meanings should be based in their “mundane,”[64] everyday usages, and claims her conception of moral sanity aligns with those conventions. Her argument, however, leads us to a picture of moral sanity that is undeniably contrary to those usages.

## IV. Sanity Reformulated

For these reasons, Wolf’s position needs some tweaking. We need a sanity-condition that does not lead to conceptually unacceptable conclusions, and more accurately describes our real-life assignments of responsibility.

[63] Wolf, pg. 57.

[64] Ibid, pg. 47.

Analogize ethics to a game. To hold your chess partner responsible for making good or bad moves, she must sufficiently understand the objective of the game, how the pieces move, etc. We would not hold someone incapable of grasping these rules responsible for doing good or bad things in the context of chess, because someone who lacked the capacity to understand the rules of chess would be 'chess-ly' insane. (This is a clunky term, but the point is made.) If your partner were unable to grasp the rules of chess and happened to make a poor move, she would not be deserving of chess-ly blame; if she happened to make a good move, she would not be deserving of chess-ly praise either. The feedback only functions if the receiver has a sound understanding of the system within which they are being blamed or praised. If the receiver does not understand the constitutive rules of the system, they are no longer operating within the system, and so we cannot evaluate them by the metric *of* the system.

So to evaluate people by a moral metric they must be capable of understanding and participating in the 'system' of morality. Wolf's sanity-condition allows us to evaluate by a moral metric only people who come to the 'right' moral conclusions, but of course we can use the metric to evaluate people who come to the 'wrong' conclusions as well – that's in large part the point of the metric itself. She thinks that if one is operating 'poorly' within the system, they cannot be judged by the standards of the system. But to be judged by the system's standards one just needs to be operating within the system in the first place. That is why moral sanity consists in the capability to understand the system itself.

The conditions under which valid moral feedback is given, then, will depend on what the 'game' of morality looks like. Offering a robust definition of morality here would exceed the scope of this paper, but the element that I think is key for my purposes is that morals exist in *codes* — codes of conduct.[65] They are principles that differentiate

[65] Gert, Bernard and Gert, Joshua (2002, rev. 2020). 'The Definition of Morality.' In *The Stanford Encyclopedia of Philosophy.*

between right and wrong behavior in a general sense, and are then applied to particular situations. Principles can be modified by other principles in certain complex situations, but they generally stand independently of any particular set of circumstances. Appeals to morally justify behavior, then, are appeals to abstract behavioral principles. Examples: 'act so as to bring about the greatest happiness for the greatest number;' 'pursue basic goods like life, knowledge, play, aesthetic pleasure, and sociability;' 'act in any given situation as the virtuous person would;' 'act only upon maxims that you can will to become universal laws.' Moral decisions are made by applying general rules like these to individual situations – they are never made arbitrarily, for they must be justifiable if questioned.

For beings to be candidates for moral feedback they must be capable of understanding this. They must be capable of recognizing the project of ethics for what it is: the task of formulating correct *principles* of action. Whereas only certain people play chess, and only for a given amount of time, everyone is always 'playing' the game of morality, for we are all constantly behaving.

The root of Wolf's difficulties is her offering too *narrow* a conception of sanity. Compare the euthanasia-attacker and defender, who both have a proper understanding of moral thought, and engage in good-faith attempts to justify *A* and *B*, with a young child. The child comes to a conclusion about what should be done simply on the basis of her emotional, one-time response to the situation – perhaps death upsets her greatly, so she says that the physician shouldn't help end the patient's life – and is therefore unequipped to grasp the nuance of moral thought. She cannot grasp the abstract prescriptive force of the moral arguments at play – perhaps she cannot understand what is meant by 'intrinsic' vs. 'instrumental' goods – and so can only justify her behavior on a moment-to-moment basis.

This tension between what one may *want* in the present moment and what they think is *right* in general is a hallmark of moral thought. No doubt George Washington didn't want to get in trouble for chopping down the cherry tree, but this impulse was overruled by the power of the general prescription that one should not lie. One's momentary desire may often align with one's values – but the capacity to recognize that, and act on the desire *because* it aligns with one's principles, and not merely because one desires it, is what makes for moral sanity.

I would therefore reformulate the sanity-condition as: *the capacity to justify one's actions by appeal to general principles of behavior*. And since values are merely general behavioral principles that prescribe the pursuit of something of value, we may say in even simpler terms that to be morally sane one must be capable of justifying her actions by appeal to values. When one is capable of thinking through which abstract *ought*-principles she subscribes to, she is morally sane, and thereby an appropriate target of moral praise and blame. So conceived, moral sanity is broad enough to leave room for both shifts in values over time and genuine moral disagreement between contemporaries. We may disagree with someone, but if her justifications have moral integrity, we tend not to label her insane. Only if she is incapable of formulating her values as principles – incapable of formulating an ethical argument – is she morally insane. This description of moral sanity both is internally consistent, and more fully captures the way we actually assign responsibility.

## V. Defending the Reformulation

It might be said that the picture of morality I have proposed is too broad, and that morality is just about doing the *right* thing, not any thing that one might be able to justify by appealing to general behavioral principles. But to think like this is to fall into the trap that defeated Wolf. When I refer to someone who 'thinks morally' I refer to someone who is *capable* of moral thought, not necessarily someone who arrives at the 'right' moral conclusion. Someone can still think through what they should do in a given situation and come to a poor conclusion via poor values and/or empirical considerations. But they are still capable of moral thought, albeit poor thought – and, per compatibilism, it is the relevant capability that I am trying to accurately describe. All I have said is that if one can justify some behavior *P* by explaining why the reasoning underlying *P* holds in other situations as well, and not just in the current situation, then they are capable of moral thought, and are therefore proper targets of moral feedback. This is not overly broad.

At times principles of action conflict, and we are hard-pressed to decide between them. We seem to have an evolutionary 'soft spot' for entertaining values that empathetically consider others' interests, since teamwork greatly aids survival prospects. Perhaps, however, there do exist some cases in which it is more correct to disregard these interests wholly in favor of one's own. The ethical egoist thinks so. And there are plenty of other points of disagreement: there are virtue ethicists, hedonistic utilitarians, preference utilitarians, deontologists, natural lawyers, new natural lawyers, feminist ethicists – the list goes on. The disagreements between these camps concern the particulars of moral theorizing and action – but regardless of the particulars of their plans of action, they all justify their plans by appealing to general principles.

It might be objected that my reformulation overly focuses on agents' capacity for principled, rational action, and fails to mention some capacity for emotional sensitivity to the suffering of others. In his paper 'The Conscience of Huckleberry Finn,' Jonathan Bennett shows how "sympathy" can act as an important counterbalancing force in people who arrive at a "bad morality"[66] purely deliberatively. In freeing Jim, Huck acts in accordance with his passions – his emotions – and against his principles. If we think Huck is a valid target of moral praise, isn't emotional sensitivity sometimes a necessary condition for responsibility?

[66] Bennett, Jonathan (1971). 'The Conscience of Huckleberry Finn.' In *Philosophy*, Vol. 49, pg. 1.

Often emotional sensitivity will make for a 'good' agent. But we are interested in the conditions necessary for responsible agency itself. And there are some agents who lack real empathy that we would hold responsible. Consider the ethical egoist, who thinks that she "morally ought to perform some action if and only if, and because, performing that action maximizes [her] self-interest."[67] Perhaps the egoist feels sympathy for others; perhaps she doesn't. Regardless, we will want to hold her responsible when it becomes clear that she has the capacity to justify and act upon volitions we find objectionable. The reasons for this are identical to the ones presented in the suicide example. Making emotional sensitivity a necessary condition for responsibility would lead us to the same problems that Wolf's sanity-condition did. We often hold emotionally insensitive people responsible insofar as they've thought through their values. The sanity-condition must be broad enough to hold responsible people acting in accordance with a variety of behavioral norms, and numerous norms eschew emotional sensitivity. Emotions constitute a unique aspect of moral thought, and play an important role in moral psychology – but I don't see them as necessary for moral *responsibility*.

There are different reasons for not exercising the relevant capacity as I have described it: some people simply don't have it, others have it but it is underdeveloped, and others still have it yet willfully do not engage it. The second category could refer to someone who has unquestioningly swallowed the values of their society and never arrived at their own normative formulations. It could also refer to an adolescent who is in the process of developing the capacity. Our actual praise- and blame-bestowing practices confirm that we treat these two cases somewhat similarly – they are cautiously deserving of *some* responsibility, but not in the robust way that a fully morally rational adult is.

Huck seems to fall into this category; his capacity for moral thought exists in some basic form, but is critically underdeveloped. In rejecting his principles he begins to

[67] Shaver, Robert (2002, rev. 2021). 'Egoism.' In *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta.

thoughtfully reflect on what to do: he mulls over the circumstances in detail and agonizes over the conflict between his "general moral principles and particular unreasoned emotional pulls."[68] But he ultimately decides that, since he will feel bad either way, going forwards he will "do whatever 'comes handiest at the time' – always acting according to the mood of the moment."[69] This is the mindset of a child, a being that is driven by whims and shys away from confronting uncomfortable difficulties through open deliberation, of someone who refuses to search for principled justifications for their behavior and so is inconsistent in the quality of their actions – someone incapable of real moral thought. Huck's mistake is his failure to revise his principles on the basis of his sympathies; if he had done that, he would be a fully responsible agent. But he lacks the ability to engage in the "abstract intellectual operations"[70] necessary to effectuate that revision, and so decides to do away with principles altogether. He is therefore a less-than-clear case; perhaps he is deserving of *some* responsibility.

As for those who have the capacity and willfully do not engage it: if your chess partner who has the capacity to understand the game makes a stupid move, you would probably still hold her chess-ly responsible. This is because, if asked, upon reflection she could provide a satisfactory explanation of why her move was poor and what a better move would have been. So in holding her responsible you would be, in a certain sense, accusing her of not living up to her potential. Some chauvinists of the 50s, to return to Wolf's example, are therefore appropriate candidates for blame, depending on their capacity to justify their sexist beliefs. Others are not. Another situation in this category might be someone who performs an immoral action under pain of death. Many philosophers of responsibility have tried to show that such a person is not responsible

[68] Bennett, pg. 4.

[69] Ibid, pg. 8.

[70] Ibid.

because her will is not free, or she could not have done otherwise, or some other reason of the like. I think this way of approaching this situation is mistaken. *If* the person in question can justify her self-preservatory actions by citing some formulation of the principle ‘one ought to, or is at least justified in, valuing the perpetuation of her own life above more trivial moral ends,’ she is a responsible agent. She is not not responsible for doing as she did – she thought through her action, and performed it – but her adherence to the principle of self-perpetuation makes it inappropriate to fully blame her for her action. In fact, someone like the egoist might even think her deserving of moral praise. On the other hand, if the person in question cannot justify their actions by appeal to such a principle, then she is not responsible.

How are we to know if someone has the relevant capacity and is not exercising it, or simply doesn’t have the capacity at all? This is an important question for all compatibilist theories, not just mine. I answer: ask the agent. If they can provide a general principle explaining why they *ought* to have behaved in the way that they did, then they are morally sane, and so an appropriate candidate for moral praise and blame. If their reasoning is incoherent, or they cannot provide any morals by which to justify their behavior, then they are not an appropriate candidate because they are morally insane.

Psychopaths are an interesting case. It is unclear if they act according to a generalized schema about what is good for themselves, like the egoist, or if they are really just impulsive (i.e., lack sane second-order volitions). The former would be a valid target of praise and blame; the latter would not. No doubt there is some variation – and, resultantly, inconsistency in the definition of psychopathy. Psychopaths do not really undercut the intuitive appeal of my conception of moral sanity/responsibility, I don’t think, because our responsibility-practices are complex. There is disagreement about how to handle some agents. Compatibilism’s ability to account for hazy cases like these is part of its appeal. The idea that the moral capacity is something that must be developed is an old one; Aristotle thought that moral character developed only over time and by a familiarity with practical ethical situations.[71] We become responsible agents as we come to a full understanding of what moral thought *is* through regular exposure to situations in which people praise and blame us, and as we become capable of critically reflecting on that praise and blame's appropriateness (its accordance with general principles of action). This development takes place most crucially throughout childhood and adolescence;[72] no doubt it continues through adulthood as well. The question of whether psychopaths experience this development seems an empirical one, and not one I am prepared to take on here.

My formulation of moral sanity is purposely broad enough to encompass *all* moral judgments. It is important to stress that in this broadening I am not endorsing moral subjectivism, nor arguing against objectivism. I am not implying that there cannot be correct or incorrect judgments; the realm in which this paper is operating is one step removed from evaluating any particular moral judgment. It is concerned with figuring out what counts as a valid moral judgment in the first place, and arguing why the capacity to properly justify these judgments is what constitutes moral sanity. The fact that one arrives at a particular judgment, correct or incorrect, is not sufficient grounds for labeling them insane. There are more relevant pieces of the puzzle.

## VI. Conclusion

As reflective creatures – creatures capable of second-order volitions – it behooves us to come to our own moral conclusions. These conclusions will often be contrary to wide intersubjective consensus, but that is not a bad thing, for exposing our beliefs to criticism (both the criticism of popular opinion, and our own) only strengthens them. To do this

[71] Homiak, Marcia (2003, rev. 2019). 'Moral Character.' In *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta.

[72] Fine, Cordelia and Kennett, Jeanette (2004). 'Mental impairment, moral understanding and criminal responsibility: Psychopathy and the purposes of punishment.' In *International Journal of Law and Psychiatry*, Vol. 27, pg. 425–443.

properly, Wolf correctly notes, we must be morally sane. Her sane deep-self view satisfactorily answers the problems that defeat Frankfurt's 'plain' deep-self view. But her formulation of sanity leads to conceptually unacceptable conclusions, and in key cases doesn't match up with our real-life responsibility-practices. In its explicit normativity, her view fails to leave adequate room for genuine moral reflection. Reformulating sanity as the capacity to engage in this reflection, I think, strengthens the sane deep-self view greatly.

## Works Cited

Wolf, Susan (1987). 'Sanity and the metaphysics of responsibility.' In *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*. Edited by Ferdinand Schoeman. Cambridge University Press.

Frankfurt, Harry (1971). 'Freedom of the Will and the Concept of a Person.' In *The Journal of Philosophy*, Vol. 68, No. 1, pg. 5–20.

Hamilton, Jill (n.d.). 'Ethics Case Studies: Using the 'Holocaust' Metaphor.' *Society of Professional Journalists*.

Gert, Bernard and Gert, Joshua (2002, rev. 2020). 'The Definition of Morality.' In *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta.

Homiak, Marcia (2003, rev. 2019). 'Moral Character.' In *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta.

Fine, Cordelia and Kennett, Jeanette (2004). 'Mental impairment, moral understanding and criminal responsibility: Psychopathy and the purposes of punishment.' In *International Journal of Law and Psychiatry*, Vol. 27, pg. 425–443.

Bennett, Jonathan (1971). 'The Conscience of Huckleberry Finn.' In *Philosophy*, Vol. 49, pg. 123–134.

Shaver, Robert (2002, rev. 2021). 'Egoism.' In *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta.