

Responsibility, Sanity and the Reactive Attitudes

Introduction

I assume moral responsibility is possible and thus deny the association of moral responsibility with the metaphysical notion of alternate possibilities or ‘free will’. The main problem that I am concerned with is *how* moral responsibility is possible, and I seek to propose an account of this according with our intuitions as to (a) what it is to be an agent, and (b) how we *utilize* the notion of responsibility. For (a), Frankfurt’s conception of the deep self (the consistency of our higher order volitions with our first order desires), and Wolf’s condition of sanity are together defensible. With (b), Strawson’s conception of the reactive attitudes is the most coherent. I finally argue that the deep-self view is supported by an account of the reactive attitudes. My view that reactive attitudes point *beyond* themselves to a conception of the agent (who has a deep self) departs from Strawson’s theory, but allows for a holistic answer to the problem of how moral responsibility is possible.

1. The Deep-Self View of Responsibility

Frankfurt has a claim that there are two kinds of freedom (freedom of action and of will) – and that freedom of will is the freedom to will whatever one wants to will, whereas freedom of action is the freedom to do what one wills to do. We do not just have first order desires, to do just what we want to do – but we have second order desires (desires about what desires to have). Furthermore, we have second order volitions – not only do we have certain second order desires, but these desires are the kind that move one to act in accordance with them – they are part of his will. To have genuine freedom of the will, we must have the capacity to form second-order volitions – this ‘*deep self*’ essentially makes us agents. We *also* require the capacity to conform one’s first order desires to these volitions to be truly free (we need to identify with our first order desires). Frankfurt gives us an example of an unwilling addict. He hates his own addiction, but has the first order desire to take narcotics. He clearly wants his desire to refrain from taking narcotics to constitute his will (what he acts in accordance with), and therefore has second-order volitions to refrain. By failing to identify with his first-order desires, he is less morally responsible than a willing addict. The willing addict shares similar first order desires to the unwilling addict but does not have the second order volitions he is free to have (to desire to desire not to take the narcotics). He is thus morally responsible for his addiction – or at least more so than the unwilling addict. The significance of Frankfurt’s argument is the idea that responsible agents are in control of themselves in some deeper sense – that they can have second order volitions and conform them to their first order desires.

2. Where and how do we locate the ‘deep self’?

An objection the deep-self view faces is that, assuming that we could have more than second order volitions, of the finite levels¹ of higher order volitions we do have, which of these do we arbitrarily pick as being the level at which we have freedom of

¹ We cannot conceive of actually having infinite degrees of higher order volitions – I discount this possibility here (the notion, related to human agents, makes no sense).

will? An individual may feel that he relates far more to his desire to desire to run a race, rather than his desire to desire to desire to run a race. Hence, Frankfurt faces the objection that different individuals' deep selves may be located at different levels of 'higher order' volitions.

This is not an objection that undermines the deep-self view. It shows that it requires a modification. The objection presupposes that the individual whose deep self (higher order desires and volitions) is in question is a reliable source for dictating the level at which his deep self, or freedom of will is located. The example of Peter Sutcliffe, who was convinced that his murder of a number of prostitutes was divinely inspired suggests otherwise. He had a first order desire to kill the prostitutes, and he still had the second order volitions to (he desired to desire to kill the prostitutes). It is plausible that Sutcliffe would accept full responsibility and freedom of will at the level of his second-order volitions. Nevertheless, a court may discover that Sutcliffe has higher order volitions that he is not aware of to desire *not* to kill the women (perhaps because he is a psychopath or has a particular mental condition). As a consequence of its awareness of this, the court's intuition would be to at least *mitigate* the degree to which Sutcliffe is morally responsible, taking into consideration his highest order desires and volitions that he himself may be unaware of. There is a case for the deep self to be located at the highest order of volitions that exists for a person independently of where he himself claims it is located.

3. The deep-self view requires a condition of sanity

Another problem for the deep-self view suggests that whilst Frankfurt's account is not an incorrect representation of how moral responsibility is possible, it is insufficient. Wolf argues that individuals must still reach a point where the deepest self is governed by something 'logically external to the self altogether'² – including values. This additional condition is motivated by cases in which we claim some individuals are not morally responsible in the sense that others are (when they do have higher-order volitions to do the same thing). Wolf gives an example of JoJo – the son of an evil, sadistic dictator. JoJo takes his father as a role model, and develops similar values to him. JoJo has certain desires he acts in accordance with, and when he commits genocide – he is not coerced in the ordinary sense: he genuinely wants to be the kind of person he is. Nevertheless, given his upbringing, it is dubious that he should be *fully* responsible for what he does – intuitively his upbringing at least mitigates his moral responsibility for his acts. Frankfurt's deep-self view does not differentiate JoJo's father, a fully responsible individual who commits genocide, from JoJo.

To differentiate such cases, it seems plausible to stipulate that an agent must be sane, in the sense that he needs to just *be* a certain way, and have certain values, desires and motivations. The desire to be sane, as characterized by Wolf, is the desire 'to be controlled by the world in certain ways and not others'³, such that one knows what they are doing is right or wrong. Therefore we do postulate an external, normative standard of

² Wolf, Susan (1987), 'Sanity and the Metaphysics of Responsibility', in F. Schoeman (ed.), *Responsibility, Character and the Emotions*, (Cambridge University Press), pp. 387

³ Ibid, pp.381

values controlled by a common conception of the world that we share (ideally an ‘accurate’ perception of the world). It is a central criterion of moral responsibility that those who are morally responsible at some level share a certain set of ethical values and motivations – that they are sane. This common set of ethical values and aims must feature within their ‘deep selves’ (their highest order volitions), for them to be regarded as *free* and therefore, morally responsible.

Objections to and defending the sanity condition

(A) On Wolf’s theory there is a sense in which our sanity (set of ethical values and aims) is not self-created, and in some prior sense, is unavoidable. This opens up an objection: the condition of sanity is tantamount to the claim that one is either moral, or not – which implies we are determined. This is problematic for the claim that there are certain values that we all *choose* to follow and believe in and thus, we can be held to be morally responsible for our decisions. Wolf’s response would be to assert that our deep selves in one sense is certainly as unavoidable as JoJo’s deep self is – we are obviously not capable of ‘creating’ our deep-selves. However, the difference between his freedom of will and ours, is that the *kind* of unavoidability JoJo’s deep self has is different, because there are features of his character that are unavoidable despite the fact that they are seriously misguided or mistaken. His ability to *know* the right action from the wrong one is absent, and thus, the ability to revise himself on the basis of this knowledge is also absent. Ordinary sane individuals are aware of the difference between right and wrong and this allows for *self-correction* on the basis of this knowledge – an ability the ‘insane’ lack. The fact that ‘sane’ individuals have the capacity to self-correct therefore implies that they *are* morally responsible, and to a greater degree than the ‘insane’.

(B) The defensibility of Wolf’s sanity condition rests on a presupposition underpinning the entire discussion. This presupposition is made apparent by her answer to a question she poses - ‘how are ‘we’ saner than non responsible individuals?’⁴ She admits that we cannot *conclusively* make the claim that we are ‘saner’ than others. She also argues the only way in which we can infer a common set of normative reasons to act is from the existence of widespread intersubjective agreement, and the way in which we socially interact and engage with others. Widespread intersubjective agreement as to certain values (‘we ought not to kill the innocent’) is most clearly manifested through the existence of our reactive attitudes – personal attitudes of gratitude and resentment, and the moral attitudes of blame and praise. If we can establish that reactive attitudes point towards sanity and the deep-self view, then we can establish that they are essential for an account of how moral responsibility is possible.

4. The reactive attitudes direct us towards a deep self and the sanity condition

a) Strawson’s Argument

Strawson’s argument for the reactive attitudes indicates how widespread intersubjective agreement as to moral responsibility operates. On this theory, it is the attitudes and reactions of individuals, including those of gratitude, resentment, forgiveness, to which we attach importance. If a rock is dropped on my head, then the only ‘pain’ I feel is the

⁴ Wolf, Susan, op cit, pp. 386

physical one. However, if there is some intentional feature of the action – someone deliberately drops a rock on my head, I feel a degree of resentment towards the individual not felt otherwise. Such feelings obviously do exist– we attach substantial importance to the attitudes of those around us, and to the attitudes we ourselves hold. Specifically moral attitudes (blame and praise) that we hold are themselves a subset of our reactive attitudes - though they are both humanly and logically analogous, is that with the latter, we are concerned with how the particular individual reacts towards an offender. With the former, we have reactions to the quality of others' wills towards others – if a person intentionally drops a rock on someone else, I will not feel the same degree of resentment towards them. I will feel a degree of moral indignation most individuals would share with me. By sharing a common set of attitudes, desires and motivations, our reactions to the offender's quality of will towards the victim bring out what is shared in our attitudes towards each other.

We argued that cases in which an individual is not to be regarded as morally responsible are not sufficiently explained by the absence of alternate possibilities. The reactive attitudes allow us to consider these situations in more depth. Occasions for waiving blame divide into two kinds of situations in which we are not expected to feel morally indignant towards the individual. The first kind is the situation in which the individual was not sufficiently informed to make the right moral decision. Such a situation is not problematic, because we can still treat the individual as sharing our attitudes and as a morally responsible agent, despite being misinformed and thus not responsible for that specific decision. We can still hold a participant's attitude to such a person – in general, he falls within the domain of our discourse about the 'reactive attitudes'. The second kind of situation is far more significant. If we are aware that the agent's picture of the world is distorted and is lacking in moral sense, he inhibits attitudes of moral disapprobation in a different way. As a result, our attitudes towards the agent can only be from an objectivised perspective, seeking to rehabilitate and deter as opposed to blaming him. We therefore cannot view him as engaging within our moral discourse of blame, praise and desert.

b) Extending Strawson's Argument – Holism

The manner in which we cannot apply reactive attitudes to some individuals, because they have a distorted conception of the world is akin to Wolf's characterisation of our treatment of the 'insane' individual. Our punishment of the 'insane' on Wolf's view also takes on an objectivised nature of treatment through rehabilitation and deterrence. It seems that we can only view individuals as falling within the domain of our discourse about the 'reactive attitudes' if they are 'sane'.

Cases of attributing complete moral responsibility to sane individuals are therefore also cases in which we take a 'reactive attitude' towards individuals in our society. We should clarify the nature of statements about these reactive attitudes. My position (*different to Strawson's*) is that they have belief-content – they make factual claims. These factual claims are most plausibly claims about our deep selves and sanity. As a result, any reactive attitude statement of the kind 'I blame Sutcliffe for the murder of my sister' expresses my belief that Sutcliffe has a deep self and is sane to the degree that I can hold him *fully* morally responsible. Reactive attitude statements may serve other

functions⁵ but they must have belief content. Whatever it is we are directing ourselves towards when we have reactive attitudes is our conception of moral responsibility. My *additional* claim is that Wolf's deep-self view and sanity condition together is a theory of that which we direct ourselves towards when we have reactive attitudes. Therefore, as I have argued, we are morally responsible when we are sane and have 'deep selves'. The existence of and our use of reactive attitudes points to and demonstrates this.

By arguing for the deep self and sanity view with the reactive attitudes, I strengthen both approaches to proffer a holistic idea of how moral responsibility is possible. If the existence of the reactive attitudes alone were an account of how moral responsibility is possible, and they did not direct us to abstract claims about the kinds of individuals we are, we would have a counterintuitive view of responsibility. We believe claims about moral responsibility are claims about agents and how agents are – we do not have indiscriminate reactive attitudes towards objects, for example.

Conclusion

I therefore contend that the sanity condition, the deep-self view and the reactive attitudes together show how responsibility is possible. The deep-self view explains how our status as agents is central, whilst the sanity condition indicates why we sometimes *mitigate* the responsibility of those who do have deep selves. Since the sanity condition commits us to an objective standard of values, these need to be justified. An appeal to reactive attitudes allows us to do so via its characterization of intersubjective agreement. The nature of reactive attitude statements themselves also directs us to claims about the deep-self view and sanity condition. My account of how responsibility is possible is thus a holistic one.

⁵ Such as expressing emotions or prescribing actions to others.