Bringing the deep self back to the racecourse: Rethinking accountability and the deep self

Ke Zhang

This is the uncorrected proof of a paper published in Analytic Philosophy.

Please cite the published version.

Abstract

Deep self views of moral responsibility suggest that an agent fully satisfies the freedom condition for responsibility if and only if her actions or omissions issue from, and so express, her deep self. This analysis generates both false negatives and false positives regarding people's responsibility, and counterexamples proliferate. I defend a novel version of the deep self view by offering a necessary condition for accountability, while retaining the core of deep self views. Indeed, an agent may be blameworthy for her wrongdoing without it issuing from, and so expressing, her deep self. And yet, I argue that she must have a deep self for which she is responsible. This is achieved by paying closer attention to history than standard views have. Focusing on history then reveals a less discussed problem for standard views: the ahistorical features of them make them less equipped to explain cases of blameworthiness that is undermined.

Introduction

Deep self views remain among the most appealing contemporary theories of moral responsibility. Deep self theorists contend that an agent's deep self plays the grounding role in virtue of which an agent acts freely and is responsible for what she does (Frankfurt, 1971; Watson, 1975, 1987; Bratman, 1997, 2004, 2005; Shoemaker, 2015; Sripada, 2016; Gorman, 2019). This is insightful. These views address the question that freedom understood in the traditional Hobbesian and Humean sense as in "lack of constraint" requires that a constraint arising from within one's own psychology be accounted for. They do so by providing resources to mark a division within an agent's self to capture her deep self—who she truly is.

With the deep self identified, when an agent acts freely her action is an expression of who she truly is. And when she acts freely, she is then a candidate for being responsible, insofar as her action or omission is an expression of her deep self.

But standard deep self views tend to imply both false positives and false negatives regarding people's responsibility. Think about Susan Wolf's well-known Jojo (Wolf, 1987), who is severely indoctrinated by his evildoing father. Growing into an evildoer himself, Jojo's morally objectionable behaviors issue from, and so express, his deep self. He thus counts as free and is responsible for his evildoing on these views. Nevertheless, given his turbulent and distorted personal history, one can reasonably question whether acting from a deep self really is sufficient for acting freely and responsibly.

Consider, again, someone who loves spicy food, but resolves to not eat spicy food for the sake of not irritating her stomach ulcer, who then freely and responsibly acts contrary to her deep self with regard to physical health, and puts too much red pepper powder in her meal out of weakness of will. One may question whether, after all, acting from a deep self is necessary for acting freely and responsibly. Despite renewed interest in deep self views (Strabbing, 2016b; Matheson, 2019; Gorman, 2022), these problems remain troubling, and deep self views have remained unpopular.

In this paper, I defend a novel version of the deep self view in which an agent must be responsible for her deep self to be accountable for what she does, where the deep self in my view is understood as crucially expressed, albeit not exhausted by, her deep valuing and deep values.[1] Responsibility for the deep self, in turn, requires a historical condition;[2] that is, for an

[1] Throughout the paper, I will discuss *value* both as noun and as active verb. This is because I believe that 1) an agent with a deep self is one who has things about her self and in the world that she deeply values, and 2) that which she values deeply are her deep values. The second claim concerns the relationship between *valuing* and *having what one values as one's values*. To further explore and defend this claim deserves a full project of its own, and I will have to leave that task to another paper. In the following discussions, I will sometimes drop either value as noun or value as active verb, depending on my emphasis on the discussion at hand.

[2] This historical condition is different from what we might call a positive one and what might call a negative one. A positive historical condition would say that what is required for responsible agency is a certain kind of history—one that involves responsibility-conferring elements or one that is free of responsibility-defeating elements. A negative historical condition would say that what is necessary for responsible agency

agent who has a personal history, she must be afforded, at some point in her history, an unimpeded opportunity to develop and exercise a certain self-constituting ability.[3]

Before I proceed, let me state two important qualifications. First, although my view advances an account of responsibility in which an agent is *accountable* for her behaviors in the sense that renders her liable to blame or praise, not all deep self theorists claim to argue for accountability. While traditional deep self theorists argue for the purchase of an agent's deep self on the grounding of accountability, more recent development of the deep self view tend to focus only on responsibility in the *attributability* sense.[4] According to these views, an agent is attributionally responsible for her behaviors just in case they are properly attributable to her deep self, which requires standing in the right relation with it. Depending on which of these views the reader has in mind, the purpose of my paper would be two-fold. The primary goal is to offer a necessary condition for accountability that avoids generating false positives and false negatives regarding people's responsibility, like traditional deep self views do. This is in order that we can explain both cases in which an agent is blameworthy for a wrongdoing that is indeed not expressive of her deep self, as well as cases in which an agent's blameworthiness is undermined even though her action is expressive of her deep self.[5] But if

is the absence of a certain kind of history—one that involves responsibility-defeating elements. A positive condition denies that agents without a personal history can be responsible agents. Here I have in mind instant agents, like McKenna's Suzie Instant (McKenna, 2004) and Mele's minutelings (Mele, 2013). A negative condition allows for both agents who have a personal history that is not of the responsibility-defeating kind, and instant agents who do not have histories at all to be candidates for responsible agency. With the historical condition I propose, I limit my attention to agents *with* personal histories, and leave it open whether agents who do not have personal histories might be responsible agents in some other way. For a thorough survey of rationales for historical theses for properly characterizing responsibility both as historical views and as opposed to non-historical views, see McKenna (2012). Thank you to an anonymous reviewer for inspiring me to make this clarification.

3 Here, I leave it as an unsettled matter how this appeal to abilities trades in the dialectic between compatibilism and incompatibilism. My understanding of abilities is permissive between general abilities and specific abilities, and I take the self-constituting ability pertinent to responsibility for one's deep self to be an ability developed and retained by the agent over an extended period of time. When circumstances and opportunities obtain, the exercise of that ability would render an agent's relevant specific ability. But this is consistent with an agent's responsibility for her deep self when she does not exercise that ability. I thank Michael McKenna for suggestion here.

4 For deep self views of accountability responsibility, see Frankfurt (1971), Watson (1975, 1987), and Bratman (1997, 2004, 2005). For deep self views of attributability, see Shoemaker (2015), Sripada (2016), Strabbing (2016b), Matheson (2019), and Gorman (2019).

5 It has been argued that conditions for blameworthiness can be met without the occurrence of any wrongdoing. For some of the relevant discussions, see Khoury (2011) and Capes (2012). It is beyond the scope of this paper to discuss instances of blameworthiness without wrongdoing. Instead, I will focus only

the reader has in mind those recent views of attributability, then the secondary goal of my paper would be to offer a deep self view of moral responsibility to account for cases in which an agent's blameworthiness for her wrongdoing, for which she is indeed attributionally responsible, is nevertheless undermined.[6] Second, I will restrict my discussion to cases of blameworthiness, and, specifically, of blameworthiness that is undermined. I leave it as an unfinished philosophical project to further explore whether the same implications can be drawn for praiseworthiness.[7]

In what follows, I will lay the groundwork for my view in section 1. In it, I introduce what I take to be the deep self, and develop two novel claims regarding a long-neglected thesis of responsibility for the deep self. Drawing upon resources from my thesis of responsibility for the deep self, I present my deep self view of moral responsibility in section 2. I then illustrate a crucial feature of my view with examples inspired by Susan Wolf's well-known case of Jojo in section 3. The examples I give will reveal the historical dimension unique to my view. In section 4, I further discuss the historical dimension and how it sets my view apart from standards views by absorbing Susan Wolf's criticism of those views, which will bring new life to the deep self view of moral responsibility.

1. Laying the Groundwork: Responsibility for the Deep Self

When we reflect on responsibility for what we do and its downstream consequences, one familiar line of thinking draws attention to how we are as persons who possess a particular

on cases of blameworthiness for wrongdoings.

[6] Achieving this secondary goal can be further assisted by arguments that aim to distinguish conditions for blameworthiness from conditions for attributability (Watson, 1996; Scanlon, 1998; Levy, 2005; Shoemaker, 2011). Although different authors have different agendas on distinguishing them, I take it as an underlying thought that judging an agent to be blameworthy concerns a broader set of facts about her than assessing her attributability does. If this is right, then recent deep self views of attributability are ill-equipped to account for an agent's blameworthiness that is undermined in cases where she is attributionally responsible.

[7] Thank you to an anonymous reviewer for suggestion here.

set of psychological traits. When our actions or omissions issue from, and so express, those traits, they are ours, and we are responsible for them.[8]

I develop a modified version of this approach: to be responsible for what we do and all that we bring about that renders us blameworthy for our wrongdoings, we must be afforded, at some point in our personal history, an unimpeded opportunity to develop and exercise a self-constituting ability to fashion our selves, and thus be responsible for our selves. Furthermore, we must also retain the ability to deploy that self-constituting ability in order to modulate our behaviors.[9] My modified approach will help develop the intuitive idea that *to be responsible for what we do, we must first be responsible for who we are.* Despite lacking explicit development, this idea has long been shared by moral responsibility theorists from different camps (Wolf, 1990, 2015; Kane, 1999; Fischer and Ravizza, 1998; and Ishtiyaque Haji, 1998). To be clear, my approach should be distinguished from the thought that responsibility for who we are must be a consequence of previous actions for which we are responsible. If that is the case, one may reasonably wonder, who is responsible for the self that authored those actions for which we are responsible? This could go on and on. Rather, I propose that certain forms of self-shaping, both active and passive, are conducive to the initial emergence of our responsible agency, and it is not the other way around.[10] As I will argue, developing an adequate thesis of this long-neglected topic on

[8] Views that take this approach most prominently consist of Harry Frankfurt's (1971) account on free will and hierarchical desires, Gary Watson's (1975, 1987) account on free agency and valuational system, and Michael Bratman's (1997, 2004, 2005) view on responsible agency and planning agency, among others. With renewed interest in this approach, philosophers such as David Shoemaker (2015) and Chandra Sripada (2016) argue for the condition for attributability in terms of an agent's cares and commitments, Jada Twedt Strabbing (2016b) offers a conjunctive sufficient and necessary condition for attributability in terms of an agent's judgments for normative reasons, Benjamin Matheson (2019) proposes an ideal narrator that connects an agent's moral identity in different person-stages that confers her attributability, and August Gorman (2019, 2022) puts forward a conjunctive sufficient and necessary condition for attributability in terms of an agent's partial and hypothetical approval for her behaviors were she to reflect on them.

[9] See McKenna (2019: 10-12) for the discussion that inspires this point. And see Strabbing (2016b: 752-754) and Strabbing (2016a: 300-305) for a similar point, where she rightly points out the importance of possessing the responsibility relevant ability, rather than merely exercising it.

[10] This claim about the initial origin of responsibility for the deep self, that appeals to self-shaping actions or omissions, requires a careful, independent treatment. Although significant, I will leave this task to another paper. For a critical discussion of the relevant worry, see Galen Strawson (1994: 6-7; 18-19) where he raises a skepticism about moral responsibility for actions that results from what he calls a paradox of moral responsibility for one's self. For proposals in response to this paradox, see Kane (1996: ch. 5) and Callard

responsibility for the deep self will reveal a far more refined relation between what we do responsibly and the deep self.

1.1 The deep self

I take the deep self to be consisting of a set of psychological features that develop and persist *over time*, and they are expressed crucially through, albeit not exhausted by, an agent's *deep valuing* and *deep values*. So, a historical dimension of the deep self and a special kind of evaluative element are significant to my understanding of the deep self. In addition, the focus on the historical dimension is relevant to this extra level of evaluation and reflection in understanding the deep self. Let me start with the evaluative element.

By deeply valuing something, I mean that an agent judges it to be good, and desires it primarily for those reasons for which she judges it to be good. In addition, she is susceptible to a range of emotions responsive to it. These three aspects of deep valuing do not merely co-occur but are connected by the reasons to which an agent is sensitive, and such a sensitivity need not be conscious, or in line with what is objectively or uncontroversially good. More importantly, that which she values in this way partly constitutes and crucially expresses her practical identity[11] in the realms of morality, society, aesthetics, and physical and mental well-being, albeit, possibly, in a disparate manner.[12]

(2008: chs. 5, 6). Thank you to an anonymous reviewer for inviting me to respond to this worry.

[11] Practical identity is a primitive of my view. I take it to be expressed by an agent's practical stances in the realms of morality, society, aesthetics, and other important realms of human life (albeit usually in a disparate way). The practical stances that a practical agent takes are explained by the multitude of her practical attitudes in the relevant spheres. Taking these practical stances makes her the practical agent she is in the relevant realms of human life. But this is not to say that a practical stance that an agent takes is exhausted by her practical attitudes. Moreover, an agent's practical identity, though crucially expressed by her practical stances, is not exhausted by them, either. For instance, an agent must be able to put in practice her practical stances to *incorporate* her deep valuings into her practical identity *not* as means to incorporate her other valuings into her practical identity. So, a practical identity is not reduced to an agent's practical stances, practical attitudes, or deep valuings. I thank Carolina Sartorio for her suggestion to clarify this point.

[12] It is a vexing matter and so I mean to leave it as an unfinished philosophical project to fully state all the elements that bear on the constitution of one's practical identity. I thank Michael McKenna for his suggestion to make this clarification.

Given how an agent may pursue and fulfill values in different ways in these different realms, an agent with a deep self does not have to have that deep self *as a whole*, and as a matter of fact, many of us do not. Instead, our deep valuings in different realms of human life express different parts of our practical identities and our deep selves. As a result, we may have a deep self in the realm of morality, but not one in the realm of aesthetics. Or, we may later develop a deep self in the realm of aesthetics but only with regard to, say, the aesthetics of food, but not regarding the aesthetics of fine art.

How deep is deep valuing? To compare, consider first Alfred Mele's characterization of valuing something in the *thin* sense, which involves a conjunction of a positive motivational element of desiring it and an evaluative element of judging it to be good (1995: 116).[13] For instance, consider Peta who desires to eat an ice cream sandwich because her friends bet that she would do it, or because it is the only thing left in the freezer and she craves snacks in the moment. Independently, she judges ice cream sandwiches to be good. However, she would rarely be motivated to go and get an ice cream sandwich without further enticement or under exceptional circumstances—like winning a bet or craving snacks with no other options besides an ice cream sandwich. She would count as thinly valuing ice cream sandwiches in Mele's sense because there is both positive motivational and evaluative components in her valuing that ice cream sandwich. However, ice cream sandwiches do not matter much to Peta and her practical identity in the realms of aesthetics of food or cuisine

[13] Mele further distinguishes between thinly valuing something that is of importance to an agent and her personal values as follows:

> "We can say that *S* at least *thinly values X* at a time if and only if at that time *S* both has a positive motivational attitude toward *X* and believes *X* to be good. Unfortunately, accepting this analysis does not settle what it is for something to be *among one's values*…Can we properly say that *X* is among a person's values if *X* is both valued by the person and of special importance to the person? No…[T]he range of personal values under consideration can be limited to things that are *valued by* valuers and are clear cases of the valuers' values" (Mele, 1995: 116).

I accept this distinction between thinly valuing something that is of importance to an agent and having it among her values. My view on deep valuing is an extension of what it is to have something among one's values. But as I stated in the previous footnote, to further develop and defend this idea is beyond the scope of this paper.

culture—it does not speak to who she is in those realms in life; she is no true gourmand and connoisseur of ice cream.[14] Deep valuing is deeper than Peta's valuing ice cream sandwiches, because the deeply valued item constitutes and expresses the valuer's practical identity in relevant realms of human life.

Or consider Peter who also desires ice cream sandwiches; but unlike Peta, he desires to eat them on a regular basis primarily for those reasons for which he judges them to be good. Peter would also count as thinly valuing ice cream sandwiches in Mele's sense, and in this case, ice cream sandwiches mean more to Peter than to Peta. But deep valuing is still deeper than that. For Peter, ice cream sandwiches do not constitute what he is in the realms of aesthetics of food or cuisine culture: like Peta, he is no true gourmand and connoisseur of ice cream, either. Indeed, to deeply value ice cream sandwiches, among other things, an agent could be a true gourmand and connoisseur of ice cream sandwiches, who judges ice cream sandwiches to be good, desires to eat and learn about them primarily for those reasons for which she judges them to be good, and is susceptible to a range of emotions in the prospect of not having access to sustain her pursuit, for example. Her deeply valuing ice cream sandwiches makes her who she is in the realms of aesthetics of food and cuisine culture.

Though deep valuing is important to the deep self, I suggest that it does not exhaust the deep self. Indeed, there might be other explicit or implicit attitudes that an agent holds that are constitutive of who she is, but are not parts of her deep valuings and deep values. So, characterizing the deep self as crucially expressed by an agent's deep valuing does not mean that there is a privileged set of psychological features that *just is* her deep self, like many deep self theorists would have us believe.[15]

Nevertheless, clarifying that there is not a privileged set of psychological features that is the deep self makes it no less important to theorize about the deep self. Understanding our

[14] This, of course, does not mean that she must not be a true gourmand and connoisseur of, say, spices. Rather, what is important to note here is that we would not know about this aspect of her practical identity with regard to the aesthetics of food and cuisine culture by her thinly valuing ice cream sandwiches.

[15] For an insightful identification of this problem for a lot of deep deep self views, see Gorman (2022).

responsible agents as beings who draw resources from their deep valuings and deep values—from who they are, to modulate their behaviors, and as beings who are able to do so, plays a critical role in understanding moral identities. Indeed, responsible agents are beings with moral identities.[16]

Beyond that, capturing the evaluative element of the deep self reveals the historical dimension of my view. No one is born with a deep self. The true gourmand and connoisseur of ice cream sandwiches, for instance, does not deeply value ice cream sandwiches from age one. An agent obtains critical aspects of her deep self as she comes to acquire certain values from which she acts.[17] The ways in which she acquires them, then, reflect the kind of personal history she has. Focusing on the kind of personal history an agent has, as I will suggest, is critical to our assessment of her responsibility *for her deep self*: was she ever afforded, at some point in her personal history, an unimpeded opportunity (which involves, among other things, a stable and healthy household growing up, access to education and affordable health care, just society) to acquire her values, develop and exercise the ability necessary for responsibility? If her personal history is one in which at no point was she afforded such things, then she is not responsible for who she is. And if she does something morally objectionable, then her blameworthiness would be undermined.

1.2 Aspiration and two kinds of deep self

As I suggested, becoming responsible for one's deep self involves various value-engagements—we fashion who we are when we are in the pursuit of acquiring, reevaluating, retaining or

[16] See McKenna & Van Schoelandt (2015: 55-59) for their unprecedented effort in advancing a hybrid view of a mesh view of moral responsibility (deep self view) and a reasons-responsive view of moral responsibility, where they argue that the resources one can draw from one's psychological mesh (and the ability to do so) plays a critical role in understanding our moral identities.

[17] Here, I take acquiring a value to be not only involving seeing something to be good or valuable, but also seeing it to be good or valuable *to the agent*. By seeing something as good to her, an agent may desire it, judge it to be good either consciously or subconsciously, feel certain emotions towards it, or all of these things together. In addition, to acquire a value is to be disposed to live up to it. Given the time and opportunities she is afforded, she acts on what she values and fulfills her values. Depending on what her deep valuings and values are, they will then be incorporated to who she is as a practical agent in the relevant realms of human life.

rejecting, and fulfilling values. When we successfully do so over time, we become a slightly, moderately, or drastically different version of ourselves. Philosophers have discussed self-shaping in different ways, either in terms of an agent's evaluation of her own motivation, her future-directed value-pursuit, or her end-setting (Frankfurt, 1971; Taylor, 1976; Schmidtz, 1994; Kane, 1996; Korsgaard, 2009; Callard, 2018).[18] These different aspects of self-shaping can be characteristically understood through the lens of *aspiration* (cf., Callard, 2018). In light of this characterization, I offer a novel distinction between an *actual* deep self and an *aspired* deep self. They differ from each other in two following respects.

First, an agent's actual deep self is crucially expressed by values that she has already acquired; they constitute what she is *now.* An aspired deep self is expressed by values that she is able to acquire, judges or deems worth acquiring,[19] but has yet to possess; they would constitute what she wants to become *in the future*.

Second, an aspired deep self is obtained through aspiration. Although it might be intuitive from how we normally use the word "aspiration" to think that it merely suggests a future-directed attitudinal change in the cognitive sense—that is, we aspire to obtain a moral value, for example, just when we aspire to obtain a further understanding of it. Nevertheless, for aspiration to indicate an active engagement with the shaping of one's self, it cannot merely involve a cognitive state. More importantly, it must involve a conative state where an

[18] For example, Charles Taylor argues that what is important to responsibility for an agent's self is her *qualitative evaluation* of her own motivation, according to which an agent evaluates her motivation in accordance with a conception of modes of life she wants to lead, and the kind of person she wants to become (Taylor, 1976). David Schmidtz proposes that not only do agents rationally *pursue* ends for their own sake (as final ends), but they can also be justified in rationally *choosing* those ends as their final ends by means of having *maieutic ends*, the latter of which are achieved by an agent coming to choose and realize certain final ends for herself (Schmidtz, 1994: 226; 231). Christine Korsgaard suggests that as rational beings, we self-constitute by choosing to act, and by actually acting, in accordance with conceptions of particular practical identities. This makes us the authors of our actions and makers of our own identities (Korsgaard, 2009: 20, 22, 24, 42). Agnes Callard has recently argued that rational valuational transformation marks an agent's own making of becoming what she wants to become, and she characterizes this transformation as aspirational. During that aspirational process, an agent acts for what Callard calls *proleptic reasons* to obtain values that she wants to obtain and has yet to fully understand. As a result, she will understand those values more fully by having acquired them (Callard, 2018: ch. 2).

[19] Here, by deeming something as worth acquiring, I mean that an agent has not yet formed a judgment (consciously or not) either that it is good, or that it is good for her. But she may notice that she has a desire to learn about it and to take it as her own, or she may have a vague impression that it is good from other people's testimony, among other things.

agent actively takes courses of action to obtain and fulfill that to which she aspires. An actual deep self, in comparison, can be obtained through the processes of either passive or active engagement with self-shaping. For instance, at least for some people, the first deep self they acquired was acquired in an unreflective or superficial manner.

1.3 Responsibility for the deep self

Now that I have laid the groundwork for what I take to be the deep self and my distinction between an actual and an aspired deep self, let me propose two claims regarding responsibility for one's deep self. The first is a sufficiency claim, the second, a distinct necessity claim. Both trade in the two kinds of deep self.

> *Sufficient* An agent is responsible for her actual deep self if it is obtained through her aspiration to transform her previous deep self to an aspired deep self, in which case the previously aspired deep self is now her actual deep self.[20]

> *Necessary* An agent is responsible for her actual deep self only if she possesses the ability to aspire to transform her actual deep self to an aspired deep self, given the time and opportunity to do so. In addition, her not exercising this ability is by her own making, not something beyond her control.

Offering a sufficient condition and a necessary condition as two separate principles has important implications. First, an agent may not have aspired to a deep self, and yet she may still be responsible for her actual deep self on the condition that she is able to so aspire, given

[20] Here, I acknowledge that there is a challenge from manipulation and brain engineering of the sort such that an agent might come to aspire as a result of such manipulation and brain engineering. In that case, it is argued that such manipulative causes can be responsibility-defeating. See Mele (1995: chs. 9, 10) for relevant discussions. These challenges will not be addressed in this paper, and for simplicity, I will leave the condition of non-manipulation implicit throughout.

the time and opportunity afforded, but does not exercise that ability through her own making. For example, an agent who obtains her first deep self in a passive and unreflective manner, and stays idle with it despite the time and opportunity given to her to examine and shape her self for better or for worse, would still be responsible for her deep self in the way identified.

Second, an agent may have the necessary ability to aspire, and is afforded the time and opportunity to do so, yet she may not satisfy any complete set of sufficient conditions. For example, suppose she is severely indoctrinated to the extent that there is no control left in her, in that case, her not exercising her ability to aspire is brought about by things beyond her control. She is then not responsible for her deep self.[21]

These two claims about responsibility for one's deep self and the conceptual space left by them will show their significance in advancing my deep self view of moral responsibility. They will provide indispensable explanatory power for an agent's responsibility, and in particular, her blameworthiness for her wrongdoing that is undermined in the cases I will focus on.

2. A New Deep Self View of Moral Responsibility

In my modified approach to understand moral responsibility in terms of the deep self, I suggested that to be responsible for what we do and all that we bring about that renders us blameworthy for our wrongdoings, our personal histories must be ones in which we were afforded, at some point, an unimpeded opportunity to develop and exercise a self-constituting ability to fashion our selves, and thus be responsible for our selves. To complete this approach with an important detail from my two claims about responsibility for the deep self —that is, the aforementioned self-constituting ability is the ability to aspire to a different

[21] A complete theory of responsibility for one's deep self would fill the gap in the latter case; however, here I only mean to argue for a more modest theory, one that advances one sufficient condition for responsibility for one's deep self, and one that advances a distinct necessary condition. I thank Michael MeKenna for his suggestion here.

deep self—I now offer a new formulation for the deep self view of moral responsibility as following.

> *NewDS* An agent acts freely and is morally responsible for what she does that renders her blameworthy for her wrongdoing only if she possesses a deep self for which she is responsible. Responsibility for her deep self, in turn, requires that she was afforded, at some point in her personal history, an unimpeded opportunity to develop and exercise an ability to aspire to a different deep self. As she acts freely and responsibly, she retains the ability to deploy such an ability to draw upon resources from her deep self to regulate her behaviors.

My formulation turns on a far more refined relationship between the deep self and responsible agency than standard deep self views do, and thus avoids implying false negatives and false positives on people's responsibility like those views do, while sustaining the explanatory power of the deep self. To explain, consider four important implications that my formulation has on understanding moral responsibility.

First, when an agent actively shapes her deep self by performing courses of action as she exercises the ability to aspire, she acts freely and is morally responsible for those actions *in virtue of* actively shaping her deep self and taking responsibility for it. This reveals a more refined sufficient condition for responsibility for what one does that appeals to the deep self —to the extent that one is actively taking actions to shape one's deep self by exercising one's ability to aspire, one is responsible for them.

Second, an agent may act freely and responsibly in performing some act *A* and she might also be responsible for her deep self in virtue of her aspirations. Nonetheless, the performing of *A* does not involve an active engagement with her self-shaping. She thus may be responsible for *A* without it being a product or expression of her actively shaping her deep

self. A weak-willed action is an example of such a free act *A.* Standard views, on the contrary, contend that any agent who performs *A* is not responsible for it.

Third, an agent may act freely and responsibly, but she has not previously actively engaged in fashioning her deep self. For example, she acquires characteristics of her first and current deep self in a passive and unreflective way, and remains idle despite the time and opportunity given to her to actively shape her self, for better or for worse. Nevertheless, she is responsible for her deep self by possessing the ability to aspire, and retaining the ability to deploy that ability to draw upon resources from her deep self to regulate her behaviors, given that she has access to do so. She thus may be, as with the last example, responsible for what she does despite the fact that it is not a product or expression of her actively shaping her deep self. Again, standard views are committed to treating such agents as not responsible for those actions or omissions.

Fourth, an agent may possess the ability to aspire but does not exercise it, but different from the last example, her not exercising it results from a responsibility-defeating condition arising through no fault of her own. In that case, she is not responsible for her deep self. And if she commits any wrongdoing that is nonetheless expressive of her deep self, her not being responsible for her deep self undermines her blameworthiness for it. In comparison, standard views would suggest that she is indeed blameworthy. This, I suggest, is a false positive implied by standard views.

Although deep self views have been widely challenged by counterexamples to both its sufficient and necessary conditions, implying false positives is a less discussed problem for them. In the next two sections, I will turn to this problem. I propose that my view can explain "cutoff" cases like the one mentioned above—cases in which an agent's blameworthiness is

undermined, whereas standard views cannot, because of our different treatments on *whether* history matters, rather than *how much* history matters.[22]

3. Jojo, Dodo, Momo

Consider Susan Wolf's Jojo (1987). Jojo grew up with his dictator father who is an evildoer. He was raised to idolize his father and grew into a person just like him—he desires evildoing, endorses these desires wholeheartedly, and he acts in accordance with values he has acquired learning from his father. According to Wolf, Jojo has a deep self expressed by a value system he has adopted. His evildoing expresses who he is.

Now, allow me to fill in more details into this example. Suppose that the original Jojo is someone like the following.

> *The OG Jojo* Despite going through powerful indoctrination, Jojo still develops an ability to aspire to a different self. Among other things, he is able to employ an internal sensitivity and an external awareness to values including those different from his own such that he is able to sense tensions among his own values, and recognize differences between his and other values. As a result, he can be prompted to retain or reject old values, acquire and pursue new ones, and live up them. However, in reality, he is blocked from having access to learn those different values, or to see tensions among his own values. This is because the indoctrination has rendered his way of living and being so fixed that he is rarely and only superficially presented with values different from his own. In addition, the indoctrination deprives him of all the relevant knowledge with which he can recognize values different from

[22] Might there be cases in which someone is *less* blameworthy, rather than non-blameworthy, than they would have been were they to have fully satisfied the historical condition? I believe there are cases like that, and that my view is compatible with the idea that moral responsibility comes in degrees; that is, one might be more or less blameworthy than another for the same wrongdoing, and the difference lies in specific features of the individuals' personal histories. Although significant, I will not pursue this topic any further in this paper. Thank you to an anonymous reviewer for suggesting me to consider this possibility.

his own *as values.* He feels content with himself and his life, and never exercises his ability to aspire to a different, less evil self.

Given the conceptual space left by *Necessary* and *Sufficient*, we can acknowledge that despite having characteristics of a deep self, Jojo is not responsible for his deep self. Despite possessing an ability to aspire to a different self, he lacks access to exercise that ability or gain relevant knowledge to do so, thus cannot exercise that ability due to lack of conditions. Given *NewDS*, Jojo's blameworthiness for his evildoing is undermined on the condition that he is not responsible for the deep self he has, even though his evildoing is indeed expressive of who he is.

Now suppose Jojo has a triplet sister, Dodo, another evildoer who has gone through the same indoctrination process, but has since been assigned to deal with affairs that involve frequent interaction and cooperation with people from different backgrounds, and thus has been directly exposed to values different from her own.

The Zen Master Dodo Like Jojo, Dodo too develops an ability to aspire to a different self. Among other things, she is able to employ an internal sensitivity and an external awareness to values including those different from her own such that she is able to sense tensions among her own values, and recognize differences between hers and other values. As a result, she can be prompted to retain or reject old values, acquire and pursue new ones, and live up them. But unlike Jojo, she has been consistently presented with different values from her own. Besides being presented with them, she is involved in activities that are interactive and communicative to the effect that she cannot avoid recognizing that there exist values different from her own, and that there are tensions between those values and her own. Indeed, she recognizes these things. Nevertheless, at no point did she ever contemplate the differences between

the two sets of values, or consider changing her heart for better or for worse. She continues her evildoing and remains zen about it.

Given *Necessary*, we can say that Dodo *is* responsible for her deep self on the condition that she possesses the ability to aspire, never exercises it, despite being afforded sufficient access to do so. Most importantly, her not exercising that ability is through her own making. Given *NewDS*, her responsibility for her deep self renders her blameworthy for her evildoing.[23]

Now turn to Jojo's and Dodo's triplet brother, Momo, yet another evildoer in the family, who has gone through the same indoctrination process, and has been assigned, alongside Dodo, to deal with affairs that involve frequent interaction and cooperation with people from different backgrounds. He thus has been directly exposed to values different from his own just like Dodo.

The Inhibited Momo Despite being involved in dealing with affairs that consistently put him in exposure to values and reasons for actions different from his own, the indoctrination has rendered Momo so fixed in his way of living and being that he can barely recognize values different from his own, or tensions among his own values or between his and others'. This is because the indoctrination has inhibited him from developing an ability to aspire. He is not able to employ an internal sensitivity and an external awareness to different values in the first place. He goes through the motions as he interacts and does business with those who are different from him. Never at any point did he question his siblings' evildoing or his own.

[23] Again, this diagnosis of Dodo would coincide with Strabbing's diagnosis of instances of attributional responsibility. According to her Judgment Responsiveness View (JRV) (2016b: 744), an agent's being attributionally responsible for her action needs not be responsive to reasons that are *correct*. But also notice that Strabbing's JRV cannot differentiate Jojo from Dodo in terms of their difference in blameworthiness.

In Momo's case, his lack of ability to aspire to a different deep self is more straightforward in that he did not get to develop it in the first place. Given *Necessary*, he is not responsible for his deep self on the condition that he is simply not able to aspire. Given *NewDS*, like Jojo, Momo's not being responsible for his deep self undermines his blameworthiness.

The skepticism raised by Wolf (1987, 1990: ch. 2) facilitated by the case of Jojo as it is originally displayed, is meant to show that merely having a deep self and acting in accordance with it is not sufficient for responsibility. Thus, the deep self view is problematic. Nevertheless, I argue that the original Jojo case would only work in favor of Wolf's criticism if we understand Jojo's way of living and being as fixed as it is for the triplet brothers Jojo and Momo, but not Dodo. Having Dodo's case specified helps to reveal important features in an agent's *personal history* relevant to her free and responsible agency overlooked by standard deep self views. Namely, having a history of powerful and thorough indoctrination is one thing, having a history of powerful and thorough indoctrination that *left an agent with no opportunity to do something to and about her self* is another. The case with Dodo shows that the former does not entail the latter. And it is the latter that matters to our assessment of a wrongdoer's blameworthiness. I turn to further explore this point now.

4. History Matters

History is crucial to my view. Given that standard deep self views are famously ahistorical, it sets a critical difference between those views and mine. To be clear, how much or whether history matters to our assessment of responsibility is a topic of an ongoing debate in work on freedom and responsibility. My goal in this section is not to offer a decisive argument for a historical account of moral responsibility, or to argue that a fully historical account is preferable to a merely history-sensitive account.[24] Rather, my more modest goal is to propose

[24] For a fundamentally ahistorical but history-sensitive account of moral responsibility, see Cyr (2023). Our goals converge in how we seek to reject the diagnosis of full-blown blameworthiness for people like Jojo and Momo. My view diverges from Cyr's in that he would deny that Jojo or Momo's blameworthiness is 1) undermined, and deny that it is 2) undermined solely due to Jojo and Momo's histories. Thank you to an

that if my readers are friends to the idea that history matters to responsibility, and are thus concerned with familiar deep self views' ahistorical features, then my deep self view need not be an enemy.

The historical dimension in my view is two-fold. First, an agent develops and sustains her deep self *over time* as she acquires her values and realizes her deep valuings. Suppose in becoming responsible for her deep self, an agent obtains her actual deep self at t_1, aims at realizing an aspired deep self (or does not exercise her ability to do so through her own making) at t_2, and successfully realizes her aspired deep self (or does not do so) at t_3. Then active engagement through aspiration (or passive engagement through her own making) connects her deep self from t_1 to t_3.

Now suppose at t_2, while the agent is able to aspire to a different deep self and is going to aspire, she undergoes covert manipulation. As a result, a new set of desires, values, cares and commitments, judgments about normative reasons, and self-governing personal policies is implanted in her. More so, it effectively dominates her mental life, which involves but are not exhausted by a change regarding how much weight she gives to the same matters. For example, she may now give zero weight to matter *B* to which she used to give above zero weight.

In my view, she then no longer aims to aspire to a pre-manipulation deep self. Instead, she now either aspires to a post-manipulation deep self that she never would have aspired to were she not manipulated, or ceases to aspire altogether.[25] Because her active engagement with shaping her deep self before manipulation is disrupted through external manipulative influences, and these influences bypass her ability to evaluate, retain, revise or reject aspects of her self drawing upon resources from her deep self at the time, then when she aspires, her active engagement can no longer connect her deep self from t_1 to t_3. In that case, she ceases to be responsible for her deep self (at least for now) in virtue of being so covertly manipulated .

anonymous reviewer for directing me to the relevant discussions.

[25] For an example like this, see Mele's Ann and Beth (Mele, 1995: 145).

Given that she is not responsible for her deep self after such manipulation, she is not responsible for actions issued from that self. If she does something morally objectionable, her blameworthiness is undermined (at least for now).

This aspect of the historical dimension in my view concerns an agent's diachronic moral and practical identity, which differentiates my view from a number of familiar deep self views that do not require diachronic identity (Frankfurt, 1971; Watson, 1975; Sripada, 2016; Gorman, 2019). On those views, what matters for moral responsibility is whether an agent identifies with certain elements in her psychological constitution, whether that results in a mesh between second and first order desires or between an agent's motivation and her values, or whether the identification is understood as less wholehearted than with full wholeheartedness. According to those views, were an agent to undergo covert manipulation described at t_2, as long as she identifies either wholeheartedly or partially with the newly implanted and dominating set of psychological elements and acts in accordance with them, she acts freely and is morally responsible for her actions.

Although there are deep self views that do account for an agent's diachronic identity, and thus share this first aspect of the historical dimension of my view, such as Bratman's cross-temporal self-governing policies that connect an agent's practical identity over time (Bratman, 1997, 2004, 2005), or Matheson's ideal narrator providing narrative explanations that confer psychological connectedness between different stages of a person (Matheson, 2019),[26] these views do not always share the second aspect of the historical dimension of my view. The second aspect further requires an agent was afforded, at some point in her personal history, an unimpeded opportunity to develop and exercise the ability to shape her deep self. In my view, Dodo has a personal history in which she was afforded such an opportunity, but

[26] Although, note that Matheson (2019: 469) argues that when emotions are involved in the ideal narrator's explanation of an agent's actions, the explanation is essentially *dispositionally* diachronic, rather than temporally diachronic, thus is not inherently historical.

Jojo and Momo were not afforded such an opportunity in their personal histories. And this is why the assessment of their blameworthiness differs.

Without an explicit endorsement of the two aspects of the historical dimension, it is not apparent on many standard deep self views that there be a difference in assessment of blameworthiness of Dodo from Jojo and Momo—as long as the triplets act from the deep self with which they identify, fully or partially, actually or hypothetically, either via a hierarchy of desires, what they value, care, or their self-governing policies, they are equally blameworthy for their evildoing.

The lack of internal resources in many standard views to account for impaired agents like Jojo and Momo whose blameworthiness is undermined may mislead us to fittingly withhold emotions towards them. This brings us back to Wolf's criticism of deep self views. As she points out, acting in accordance with one's deep self might not be sufficient for responsibility, because whether one is responsible for one's deep self factors into our judgment of their responsibility for what they do that is expressive of that self, as quoted below.

> [W]e sometimes do question the responsibility of a fully developed agent even when she acts in a way that is clearly attributable to her real self. For we sometimes have reason to question an agent's responsibility for her real self. That is, we may think it is not the agent's fault that she is the person she is—in other words, we may think it is not her fault that she has, not just the desires, but also the values she does (Wolf, 1990: 37).

Some features in an agent's deep self may indicate her non-responsibility. For instance, Wolf argues elsewhere that responsibility for what one does issuing from one's deep self requires *sanity*. As she defines it, sanity is understood "as the minimally sufficient ability cognitively

and normatively to recognize and appreciate the world for what it is" (Wolf, 1987: 56). Further, sanity enables an agent to know the difference between right and wrong, and to correct her behaviors and improve herself accordingly (60). An agent is not morally responsible for what she does if she is unable to do so—that is, if she is insane. So, merely acting in accordance with her deep self is not sufficient for her responsibility for her actions.[27] It is further required that the deep self from which she acts be sane.

It is not explicit in Wolf's view that sanity is required by responsibility for one's actions or omissions *because* it is further required by one's responsibility for one's deep self. Nonetheless, I argue that the quote above strongly suggests that it is so: sanity indicates that one has developed the necessary ability to be responsible for one's deep self.[28] In my view, the necessary ability is the ability to aspire to a different deep self. Given this, I now suggest that my view can accommodate the sanity requirement on the condition that it be understood in a particular reading.

The sanity requirement says that an agent with a deep self is sane only if she is able to know the difference between objective right and wrong in the world. Knowing the difference between right and wrong can be understood in the following two readings. In the stronger reading, knowing the difference amounts to understanding and appreciating the right *as* right, and the wrong *as* wrong. Such an understanding and appreciation could potentially lead an agent to act accordingly. In the weaker reading, knowing the difference merely amounts to realizing that there *is* a difference between right and wrong, without involving the further understanding and appreciation of the right *as* right, the wrong *as* wrong.

It is unclear that this distinction is implied in Wolf's sanity requirement. For, to the extent that Wolf's original Jojo is concerned, the sanity requirement can simply show that he is not responsible for his evildoing because he is unable to know right and wrong in the stronger reading—he does not know right as right, wrong as wrong. However, to make sense

[27] See both Watson (1996: 240) and Scanlon (1998: 192, 279) for a similar idea.

[28] To be sure, this does not mean that Wolf herself commits to a history-sensitive theory of responsibility.

of Jojo's and Momo's undermined blameworthiness and Dodo's full-blown blameworthiness, this distinction is important. Their difference does not lie in their ability to know in the stronger sense—none of them know right as right, wrong as wrong. Rather, the difference lies in their ability to know in the weaker sense—only Dodo knows in the weaker sense. Recall that she is able to employ her internal sensitivity and external awareness to her values and values different from hers. In addition, she is afforded access to values different from her own, accompanied by successful proceedings of her social and work life. That success has rendered her a recognition that there is such a difference between what she deems right and what others deem right. And yet, she does not truly understand or appreciate the difference as an indication that her evildoing is wrong.

If we want to capture the difference in blameworthiness between the triplets by the sanity requirement, then it cannot merely mean the ability to know right and wrong in the stronger reading, as someone like Dodo is able to know in the weaker reading without knowing in the stronger reading. I thus suggest that responsibility requires the ability to know right and wrong *in the weaker reading*.

The ability to aspire critical to my view involves sanity in the weaker reading, but not in the stronger reading. This is because one does not need to know the objective right as right and objective wrong as wrong to aspire to a different self, but one does need to possess at least some moral knowledge in the weaker sense to aspire in our moral life. *NewDS* absorbs Wolf's criticism of familiar deep self views and accommodates Wolf's sanity requirement to account for what is necessary for accountability. All of this is achieved by focusing on features in an agent's personal history where the necessary ability is developed and exercised given access to relevant opportunities.

To conclude, I have developed a novel deep self view of moral responsibility in which an agent's responsibility for her deep self is necessary for her being accountable for what she does, and, more specifically, in the cases I have focused on, her blameworthiness for a

wrongdoing that is either expressive or not expressive of that self. My account retains the familiar core of standard deep self views in which an agent's deep self is essential to our understanding of her moral responsibility. Beyond that, my view has the internal resources to account for cases in which we sometimes deem it justified to withhold from blaming an impaired agent. This has been achieved by paying closer attention to the historical dimension of the deep self than standard deep self views have, which I hope has brought new life to the deep self view.

Acknowledgment

I am grateful to Rhys Borchert, Luke Golemon, Tim Kearl, Andrew Lichter, Michael McKenna, Alex Motchoulski, Carolina Sartorio, Robert H. Wallace, Sean Whitton, Yili Zhou and an anonymous referee for helpful discussions, comments, and suggestions on this paper.

Bibliography

Bratman, Michael E. (1997). Responsibility and planning. *The Journal of Ethics 1 (1)*:27-43.

Bratman, Michael E. (2004). Three Theories of Self-Governance. *Philosophical Topics* 32 *(1/2)*:21-46.

Bratman, M. (2005). Planning agency, autonomous agency. *Personal autonomy*, 33-57.

Callard, Agnes (2018). *Aspiration: The Agency of Becoming*. Oxford University Press USA.

Capes, Justin A. (2012). Blameworthiness without wrongdoing. *Pacific Philosophical Quarterly 93 (3)*:417-437.

Cyr, Taylor W. (2023). Why history matters for moral responsibility: Evaluating history-sensitive structuralism. *Philosophical Issues 33 (1)*:58-69.

Fischer, John Martin & Ravizza, Mark (1998). *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge University Press. Edited by Mark Ravizza.

Frankfurt, Harry G. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy 68 (1)*:5-20.

Gorman, August (2019). The Minimal Approval View of Attributability. In David Shoemaker (ed.), *Oxford Studies in Agency and Responsibility 6*. Oxford University Press.

Gorman, August (2022). Demystifying the deep self View. *Journal of Moral Philosophy 19 (4)*:390-414.

Haji, I. (1998). *Moral Appraisability: Puzzles, proposals, and perplexities*. Oxford University Press.

Kane, Robert (1996). The Significance of Free Will. Oxford University Press USA.

Khoury, Andrew C. (2011). Blameworthiness and Wrongness. *Journal of Value Inquiry 45 (2)*:135-146.

Korsgaard, Christine (2009). *Self-Constitution: Agency, Identity, and Integrity*. Oxford University Press.

Levy, Neil (2005). The Good, the Bad, and the Blameworthy. *Journal of Ethics and Social Philosophy 1 (2)*:1-16.

Matheson, B. (2019). Towards a structural ownership condition on moral responsibility. *Canadian Journal of Philosophy, 49(4)*, 458-480. doi:10.1080/00455091.2018.1480853

McKenna, M. (2004). Responsibility and globally manipulated agents. *Philosophical Topics, 32(1-2)*, 169–192.

McKenna, Michael (2012). Moral Responsibility, Manipulation Arguments, and History: Assessing the Resilience of Nonhistorical Compatibilism. *The Journal of Ethics 16 (2)*:145-174.

McKenna, M., & Van Schoelandt, C. (2015). Crossing a Mesh Theory with a Reasons-Responsive Theory: Unholy Spawn of an Impending Apocalypse or Love Child of a

New Dawn?. In *Agency, freedom, and moral responsibility* (pp. 44-64). Palgrave Macmillan, London.

McKenna, M., Coates, D. J., & Tognazzini, N. (2019). Watsonian compatibilism. *Oxford studies in agency and responsibility, 5*, 5-37.

Mele, A. R. (1995). *Autonomous agents: From self-control to autonomy*. Oxford University Press on Demand.

Mele, Alfred R. (2013). Moral Responsibility, Manipulation, and Minutelings. *The Journal of Ethics 17 (3)*:153-166.

Scanlon, Thomas (1998). *What We Owe to Each Other*. Belknap Press of Harvard University Press.

Schmidtz, David (1994). Choosing ends. *Ethics 104 (2)*:226-251.

Shoemaker, David (2011). Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility. *Ethics 121 (3)*:602-632.

Shoemaker, D. (2015). Ecumenical attributability. *The nature of moral responsibility: New essays*, 115-40.

Smith, A. M. (2005). Responsibility for attitudes: Activity and passivity in mental life. *Ethics*, *115(2)*, 236-271.

Sripada, Chandra Sekhar (2010). The deep self model and asymmetries in folk judgments about intentional action. *Philosophical Studies 151 (2)*:159-176.

Sripada, Chandra (2016). Self-expression: a deep self theory of moral responsibility. *Philosophical Studies 173 (5)*:1203-1232.

Strabbing, J. T. (2016a). Attributability, weakness of will, and the importance of just having the capacity. *Philosophical Studies, 173(2)*, 289-307.

Strabbing, J. T. (2016b). Responsibility and judgment. *Philosophy and Phenomenological Research, 92(3)*, 736-760.

Taylor, C. (1976). Responsibility for self. *The identities of persons*, 281-99.

Watson, Gary (1975). Free agency. *Journal of Philosophy 72 (April)*:205-20.

Watson, Gary (1996). Two Faces of Responsibility. *Philosophical Topics 24 (2)*:227-248.

Watson, Gary (1987). Free action and free will. *Mind 96 (April)*:154-72.

Wolf, Susan (1987). Sanity and the Metaphysics of Responsibility. In Ferdinand David Schoeman (ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*. Cambridge University Press. pp. 46-62.

Wolf, Susan (1990). *Freedom Within Reason*. Oxford University Press USA.

Wolf, Susan (2015). Character and Responsibility. *Journal of Philosophy 112 (7)*:356-372.