A common intuition is that manipulation undermines moral responsibility. But on what ground does manipulation exempt the manipulated agent from being morally responsible? Moreover, “[w]hat can we learn about the nature of moral responsibility for actions from reflection on thought experiments featuring manipulation” (Mele 2019, 2), and what are conditions on which we can hold an agent responsible? To answer those questions, in his latest book *Manipulated Agents: A Window to Moral Responsibility* (OUP 2019), Alfred Mele suggests a proposal concentrating on one agent’s system of value—by ‘value’ he means “[an agent] thinly values X at a time if and only if at that time [the agent] both has a positive motivational attitude toward X and believes X to be good” (Mele 1995, 116), where ‘positive attitudes’ are construed as desires (Mele 2019, 15). Mele’s account (i.e., DMR) provides four negative constrains on moral responsibility, and it goes like this:

*DMR.* If an agent is directly morally responsible for A-ing, then the following is false:

(1) for years and until manipulators got their hands on him, his system of values was such as to preclude his acquiring even a desire to perform an action of type A, much less an intention to perform an action of that type;
(2) he was morally responsible for having a long-standing system of values with that property;

(3) by means of very recent manipulation to which he did not consent and for which he is not morally responsible, his system of values was suddenly and radically transformed in such a way as to render A-ing attractive to him during t; and

(4) the transformation ensures either (a) that although he is able during t intentionally to do otherwise than A during t, the only values that contribute to that ability are products of the very recent manipulation and are radically unlike any of his erased values (in content or in strength) or (b) that, owing to his new values, he has at least a Luther-style “inability” during t intentionally to do otherwise than A during t.

(Mele 2019, 127-128)

It seems that one agent’s system of value plays a vital role in Mele’s account. One may wonder why is system of value so important? More particularly, with respect to condition (3), why can a radical and sudden altered system of value help conduct a “not-responsible” verdict of the agent? I tentatively argue that, for the purpose of this review, the best explanation thereof should be one agent’s system of value is grounded in her personal identity. Given this grounding relation, which implicitly supports Mele’s account, I point out Mele’s DMR probably entails incoherence.

To prepare us for a closer look on Mele’s emphasis on system of value, I primarily introduce a case of radical reversal called One Bad Day. It goes as follows (Mele 2019, 20-21): “Sally […] was one of the kindest, gentlest people on Earth. … Sally’s character was such that intentionally doing anyone serious bodily harm definitely was not an option for her. … Moreover, she was morally responsible […] for having the character
she had. … [One day] while Sally slept, a team of psychologists [...] implanted [vicious] values in Sally after erasing her competing values. They did this while leaving her memory intact, which helps account for her surprise. Sally reflects on her new desire. Among other things, she judges, rightly, that it is utterly in line with her system of values. She also judges that she finally sees the light about morality—that it is a system designed for and by weaklings. Upon reflection, Sally ‘has no reservations about’ her desire to kill George and ‘is wholeheartedly behind it’ (Frankfurt 2002, 27). Furthermore, the desire is ‘well integrated into [her] general psychic condition’ (Frankfurt 2002, 27). Seeing nothing that she regards as a good reason to refrain from stalking and killing George, provided that she can get away with it, Sally devises a plan for killing him; and she executes it. ... Her current view of things is utterly predictable, given the contents of the values that ultimately ground her reflection, and her new system of values left no room for mercy. Sally ‘identifies [herself] with the springs of [her] action’ (Frankfurt 1988, 54), and she kills George ‘because [she] wants to do it’ (Frankfurt 2002, 27).”

People generally share the intuition that Sally is not responsible for killing George in One Bad Day. Nevertheless, we wonder “[if] Sally is not morally responsible for killing George …, why not?” (Mele 2019, 26) Mele’s answer is probably that, given DMR, Sally is not responsible because killing George is an action from a radically and suddenly revised system of value, for the possession of which Sally is not responsible. Vargas diagnoses that pre-Sally (i.e., Sally before brainwashing) being not responsible for killing George is arguably a trivial truth—since pre-Sally and post-Sally (i.e., Sally after brain-washing) are arguably distinct persons, and we usually do not hold one
person responsible for another person’s deeds (Vargas 2013, 300). But Mele seemingly tries to avoid such involvement in debates of personal identity, in reply to Vargas he writes, “[the] path I choose is the easiest one to travel; it is to say that [Sally]—that person—persists through the value changes. Persons are one kind of thing, and moral personalities, moral characters, and practical identities are another (or others).” (Mele 2019, 75) It appears that, as Mele believes, a theory featuring moral character (like his DMR) already suffices to deal with our “not responsible” intuition in cases like One Bad Day. Back to post-Sally, nonetheless, Matheson asks that intuitively “since post-manipulation [Sally] certainly acts like a nasty piece of work in accordance with her will, why not treat her like a nasty piece of work?” (Matheson 2014, 330) Call it Matheson’s question. Mele assumably has to reject the idea of holding post-Sally responsible for killing George, because according to DMR, only “by means of very recent manipulation to which [s]he did not consent and for which [s]he is not morally responsible, [her] system of values was suddenly and radically transformed in such a way as to render [killing George] attractive to her” (Mele 2019, 127-128). To put the dispute on the table, Matheson and Mele disagree on whether we can treat the implanted system of value as post-Sally’s, and more importantly, why or why not.

Arguably, Mele’s position is that the implanted system of value, a product of radical and sudden transformation, can’t be treated as post-Sally’s. But why? It seems the radical and sudden transformation per se is not enough to ground Mele’s position. Consider the case of regeneration for illustration. Chuck, a person valuing vicious deeds, could assumably refresh himself by a radical and sudden rewriting of his system of
value, turning out to be a good person. Intuitively, we believe the refreshed system of value can be treated as Chuck’s. Thus, given the insufficiency of the radical and sudden transformation per se as grounds for Mele’s position, on what ground can’t we treat the implanted system of value as post-Sally’s then? I believe a better explanation—and as far as I am concerned, the best explanation—is that post-Sally does not self-identify as a killer, a person equipped with a system of value leaving no room for mercy. To put it another way, the implanted system of value does not represent Sally identity because were Sally not brainwashed, she would resist to be treated as a murderous person, and she would deny that she was blood-lust by nature. In a word, one agent’s system of value is grounded in her personal identity. Call it the Grounding Explanation. The Grounding Explanation looks more plausible, on the basis of which we can have a better understanding of why system of value matters so much in Mele’s DMR—since our identification, viz. our answer to the question who we really are, play a significant role in our moral judgement. We may assume Mele adopts this explanation. Back to the case of One Bad Day, Mele could defend his position in this way—post-Sally would resist to be identified as a killer only by virtue of her self-identification as a kind person, which originates to the very beginning of the story, when (pre-)Sally remains one of the gentlest people on Earth. In that sense, to apply the Grounding Explanation, Mele arguably has to presuppose the persistent identity from pre-Sally to post-Sally—not as a casual choice (as suggested in Mele 2019, 75), but as an essential precondition.

If my diagnosis about Mele’s presupposition of persistent identity is correct, then Matheson’s question, no longer a merely intuitive concern, would pose a substantial
threat that Mele’s DMR is probably incoherent. The incoherence can be illustrated as follows. On the one hand, to account for his “not responsible” verdict of post-Sally, Mele has to ground his explanation in the persistent identity of Sally throughout brain-washing. On the other hand, provided a certain thesis of personal identity (e.g., the psychological account in Matheson 2014), pre-Sally and post-Sally are distinct persons with distinct identities, i.e., there is no persistent identity of Sally throughout brain-washing. Therefore, Mele’s DMR is potentially incoherent. In spite of my argument above, I agree that Mele could reject Matheson’s assumption that pre-Sally and post-Sally are distinct persons with distinct identities. But a rejection such like should be grounded in a more specific conception of personal identity, which suggests Mele can no longer cautiously avoid an involvement in personal identity debates (Mele 2019, 75). In fact, as it seems to me, Mele’s belief, that radical and sudden transformation may undermine moral responsibility, already makes some implicit presumption thereof, and my review could further push Mele to make his underpinned notion of personal identity in DMR more explicit. Summing up my point, Mele’s emphasis on system of value is arguably grounded in conceptions of personal identity, and Mele has to spare more effort developing his notions of personal identity behind DMR, which I personally believe to be a direction worth pursuing from Mele’s accomplishment in *Manipulated Agents: A Window to Moral Responsibility* (OUP 2019).

References

Harry Frankfurt. (1988). *The Importance of What We Care About*. Cambridge:


