Review of Alfred Mele's *Manipulated Agents: A Window to Moral Responsibility*

Taylor W. Cyr

Forthcoming in *Philosophical Quarterly*; please cite published version.

Alfred R. Mele, *Manipulated Agents: A Window to Moral Responsibility* (Oxford: Oxford University Press, 2019), 184 pages. ISBN: 9780190927967 (hbk.). Hardback: $65.00.

In his most recent book, *Manipulated Agents: A Window to Moral Responsibility*, Alfred Mele defends the view that an agent is morally responsible for an action only if the agent was not manipulated (in a sense I will soon illustrate) into performing that action. Consider the case of Sally from *One Bad Day*, a prominent case of manipulation in Mele's book:

> When Sally crawled into bed last night, she was one of the kindest, gentlest people on Earth. She was not always that way, however. When she was a teenager, Sally came to view herself, with some justification, as self-centered, petty, and somewhat cruel. She worked hard to improve her character, and she succeeded. When she dozed off, Sally's character was such that intentionally doing anyone serious bodily harm definitely was not an option for her: Her character—or collection of values—left no place for a desire to do such a thing to take root. Moreover, she was morally responsible, at least to a significant extent, for having the character she had. But Sally awakes with a desire to stalk and kill a neighbor, George. Although she had always found George unpleasant, she is very surprised by this desire. What happened is that, while Sally slept, a team of psychologists…implanted [new] 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…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—and him—that afternoon…When Sally falls asleep at the end of her horrible day, the manipulators undo everything they had done to her. (pp. 20-21)

A lot of recent work on the nature of moral responsibility is informed by intuitions about cases like *One Bad Day*. Mele judges that Sally is not morally responsible for killing George (and, as Mele reports in a brief appendix, the majority of participants in some experimental philosophy about similar cases seem to agree).

Before we consider why Sally seems not morally responsible for killing George, note that it is possible for Sally to satisfy any purely *internalist* (or *structuralist*) conditions on moral responsibility. These are conditions that an agent may satisfy at the time of her action and make no reference to the agent's history. Consider Harry Frankfurt's influential view. According to Frankfurt, as long as an agent has no reservations about a desire to act, is wholeheartedly behind the desire, and the desire is well integrated into her general psychic condition, then the agent is morally responsible for acting on this desire. (Mele builds into *One Bad Day* that Sally meets these and other conditions—see pp. 20-21.)

Now, why should we think that Sally is not morally responsible for killing George, despite satisfying all of these internalist conditions on moral responsibility? Mele offers the following "radical reversal suggestion" (p. 26): "Sally's pre-transformation character was sufficiently good that killing George was not even an option for her; and the combination of this fact with the fact that Sally was morally responsible (to some significant extent) for that character, facts about her history that account for her moral responsibility for that character, facts

about her post-manipulation values and associated abilities, and the facts that account for her killing George suffices for her not being morally responsible for killing him." If this suggestion is correct, then an agent's moral responsibility for an action depends on her not having a certain sort of history, and thus the view that Mele defends, motivated by this idea, counts as an *externalist* (or *historicist*) view of moral responsibility.

In chapter one, Mele introduces his subject and some of the terminology to be used throughout the book. Mele rejects the view that an agent's history is *never* relevant to her moral responsibility ("unconditional internalism") because of the possibility of *indirect* moral responsibility—cases in which an agent is morally responsible for an action but only in virtue of her moral responsibility for earlier actions. The classic case here, which Mele discusses (pp. 7-8), is the case of drunk driving. Even if Van is so intoxicated that he does not realize he is impaired and unwittingly kills a pedestrian with his vehicle, Van is morally responsible for killing the pedestrian, provided that he was morally responsible for becoming so intoxicated. Mele also considers the view ("unconditional externalism") that there is no possible internal condition that an agent could be in such that she would be morally responsible for acting from that condition no matter her history. But suppose an agent, Mabel, has the ability to immediately undo any mental change brought about by manipulation (see Mele's discussion of Mabel in chapter two, pp. 15-16); with such powers, it is plausible that Mabel would be morally responsible for her actions no matter her history. Mele is surely right about both of these two types of views; however, since the internalist/externalist debate is one about direct moral responsibility, and since the debate is about the moral responsibility agent's relevantly like us (who lack the powers Mabel possesses), my own view is that this terminology is unhelpful even though Mele is right to point out that such extreme positions are untenable. In what follows, I use internalism and externalism in the standard way, where the former says that an agent's history is irrelevant to whether or not she is morally responsible for an action and the latter says that an agent's history *can* affect whether or not she is morally responsible for an action.

In chapter two, Mele discusses several cases of manipulation and uses them to challenge internalism (focusing specifically on Frankfurt's view). Mele continues this project in chapter three, addressing responses from critics, especially a series of papers by Michael McKenna. Mele begins chapter three with some background about the terminology he has used in the past, admirably noting that his exchanges with McKenna "have not quite clicked" and that he "is partly responsible for that" (p. 62) because of the way that his terminology has led to confusion. Partly because of this confusion, readers of Mele's work should take care to note that Mele has updated his own proposed compatibilist set of sufficient conditions for moral responsibility, omitting the notion of "unsheddable values" (see pp. 66-68).

At this point, one might wonder why anyone would accept internalism and deny that manipulated agents are not morally responsible for their actions. Interestingly, several compatibilists (including Frankfurt) have made claims to the effect that compatibilism requires internalism, and Mele's aim in chapter four is to show that compatibilism does not require internalism. Mele distinguishes between cases of manipulation of the "radical reversal" type illustrated above (in *One Bad Day*) and cases of "original design," such as the story featured in Mele's now famous "zygote argument" (see pp. 83ff). Mele argues that even if compatibilism requires admitting that designed agents are morally responsible, it does not follow that compatibilism requires admitting that the victims of radical reversals are morally responsible. Chapter five continues the discussion of radical reversals and original designs but focuses on the

question of when it is (and when it isn't) acceptable to "bite the bullet" about cases of manipulation (admitting the agent's moral responsibility).

Finally, in chapter six, Mele ties together some loose ends, partly by providing answers to a series of 12 questions that readers may have (see part 4, pp. 133-141). Most noteworthy from this chapter is the following: although the focus of the book has been on compatibilist accounts of moral responsibility, Mele shows that his work on manipulation is relevant for incompatibilist theories of moral responsibility too.

While readers of Mele's work (especially his recent articles on manipulation) will encounter familiar material in this book, there is some new material as well, especially at the beginning and the end, and I found it helpful to have Mele's work on the subject of manipulation collected together and organized as it is here. For these reasons, this book is a must-read for philosophers working on moral responsibility. And for those who are interested in but unfamiliar with Mele's work on manipulation, this book will simultaneously introduce Mele's previous work and bring readers up to speed on the current state of the debate.

Taylor W. Cyr
Samford University
taylor.w.cyr@gmail.com