Alfred R. Mele, Manipulated Agents: A Window into Moral Responsibility, (New York: Oxford University Press, 2019), 174 pages. ISBN: 9780190927967 (hbk.). Hardback: £ 41.99

Alfred Mele's new book explores the question of whether and how an agent's history is relevant to her moral responsibility for a present action.

Consider Chuck and Sally. Suppose that both persons are morally responsible at least sometimes for their actions and are morally responsible to some significant extent for their character. Chuck has very bad character that strongly inclines him to murder, whereas Sally's very good character makes murder unthinkable. But those treacherous neuroscientists are at it again. They erase Sally's good values and replace them with bad values that are qualitatively identical to Chuck's values. On this bad day, Sally acts solely on those new values and murders in a way that is qualitatively identical to the way in which Chuck murders; when the day is over, the neuroscientists return Sally to her pre-manipulated values. By hypothesis, Chuck is morally responsible for murder. But is Sally also morally responsible for murder? Mele's intuition is that Sally is not. But since the values, deliberations, intentions, and actions of Chuck and Sally are qualitatively identical with respect to the murder, this difference in moral responsibility evaluation must be explained by historical differences.

Mele's thesis is roughly that both compatibilists and incompatibilists should adopt a historical condition on moral responsibility to accommodate intuitions such as the one that Sally is not morally responsible for murder.

In chapter one, Mele defines the standard terms in the free will debate, and provides a brief argument for one way in which an agent's history is relevant to her moral responsibility. Consider another pair of persons, Van and Ike. Van gets himself drunk and drives home, whereas Ike is force-fed alcohol and is placed into his vehicle to drive home. Both are too drunk to know that they are impaired, and both unwittingly kill a pedestrian. Plausibly, Van is morally responsible for killing the pedestrian, and Ike is not. But because neither person satisfies the conditions on moral responsibility at the time of drunk driving, Van must be *indirectly* morally responsible—that is, his moral responsibility must be wholly derived from his previous free choices to consume alcohol in the way that he did (p. 11). Thus, a differential moral responsibility assessment of Van and Ike is grounded in their different histories.

In chapter two, Mele defends his primary contention that history is relevant to *direct* moral responsibility, which is the kind of moral responsibility that is not entirely derived from other things (p. 11). He does so in part by arguing against Harry Frankfurt's hierarchical view of moral responsibility. On Frankfurt's view, an agent is morally responsible for an action if the agent acts on a desire in a wholehearted way, but how she came to have that desire and psychic integration is irrelevant to her moral responsibility. Frankfurt's view is problematic precisely because Sally's desires can be manipulated such that she desires to kill her neighbor in a wholehearted way; and yet, we still have the intuition that Sally is not morally responsible for the murder, even though Frankfurt's sufficient condition on moral responsibility has been satisfied. (This counterexample applies *mutatis mutandis* to other history-insensitive compatibilist views.)

But why do Mele and others have this non-responsibility intuition about Sally's murder (see the book's appendix for experimental philosophical data)? Here is a Galen Strawson-style explanation: an agent must be morally responsible for the character that explains her action to be morally responsible for the action and Sally is not morally responsible for the manipulated character that explains her action. In my view, Mele rightly rejects that explanation (pp. 30-34). Instead, Mele offers "considerations that loom large" for him in explaining the intuition: "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" (p. 26, italics in the original). Mele appeals to this fact-list explanation throughout the book (pp. 37, 51, 57-58, 88, 125).

Now, the fact-list explanation does provide a window into moral responsibility. But the window would have been bigger and the glass would have been clearer if Mele were to have provided some uniting explanation about why these facts suffice to undermine moral responsibility. Such an explanation would, for example, provide insight into cases in which Sally's character is less good or she is less morally responsible for her character. It may even offer new understanding into how we become morally responsible agents in the first place.

In chapter three, Mele defends taking seriously his non-responsibility intuition about Sally's murder in response to contrary intuition pumps by Michael McKenna. Additionally, Mele argues that the non-responsibility intuition should be preserved even in Manuel Vargas's revisionist account of moral responsibility.

In chapter four, Mele argues against the common idea (à la Frankfurt, Richard Double, Gary Watson) that being a compatibilist requires affirming a history-insensitive view of moral responsibility. The common idea is that if the historical manipulation in Sally's case undermines her moral responsibility for murder, then an action's being causally determined by factors outside of its agent's control also undermines moral responsibility for the action, which amounts to the denial of compatibilism. In response, Mele denies the conditional. He differentiates Sally's case from a case of mundane causal determinism, and even from a case of "original design" in which a goddess configures a zygote and its consequent environment to ensure that the agent performs a particular action thirty years later. What is the relevant difference? It is that the aforementioned fact-list explanation applies to Sally's case but not the others (pp. 88-89). Mele concludes that compatibilism does not entail that moral responsibility is ahistorical, or, more modestly, that compatibilists have not provided a good reason to believe that such an entailment holds. But then, there is room for compatibilists to tack on Mele's historical condition to their other conditions for moral responsibility.

In chapter five, Mele continues to reflect on radical reversal and original design scenarios, and he considers the question of when compatibilists should bite the bullet on manipulation and original design cases used to support incompatibilism.

In chapter six, Mele contends that even incompatibilists should adopt his historical condition (p. 126). For although the occurrence of indeterminism at the moment of choice precludes an indeterministic agent's being manipulated to choose a particular option, the range of live options is subject to manipulation. So, if Sally is manipulated to have Chuck-style values such that the range of her alternative possibilities at a particular time is exhausted by which innocent person to kill, Mele intuits that Sally is not morally responsible for murder. This chapter also features an interesting Q&A about Mele's methodology in the book.

Mele's load-bearing argument for his historical condition rests on and only on his intuition in radical reversal cases and the accompanying fact-list explanation. Mele explicitly states that he does not rely on his appendix's experimental data to support his conclusion, but he would have been worried if it revealed that a majority of people do not share his intuition (pp. 147-148). Mele also considers other kinds of case (instant agents, minutelings, etc.), but does not appeal to them to support his position either (pp. 32, 60). One result of the exclusive focus on radical reversal cases is that his historical condition on direct moral responsibility is tailor-made to apply to only that kind of case (see pp. 66-67). As Mele says at one point, "This book takes us part of the way" (p. 38). There is still more work to be done to provide a full account of the historical condition on direct moral responsibility. Even so, the argument in this book has significant implications for ahistorical compatibilist accounts of moral responsibility, and for incompatibilist accounts of moral responsibility too. Proponents of such accounts would do well to reflect on Mele's new book.

Readers of Mele's *Autonomous Agents, Free Will and Luck*, or recent articles on manipulation will encounter familiar material in *Manipulated Agents*. But they will also notice some refinements. For example, Mele now states his historical condition on moral responsibility without reference to "unsheddable values," which is a term of art that has caused some confusion (pp. 66-68). Readers who are joining the conversation for the first time will find that *Manipulated Agents* provides an accessible overview of his earlier work as well as the state of his current thought.

Robert J. Hartman
Stockholm University
roberthartman122@gmail.com