Alfred R. Mele

Manipulated Agents. A Window to Moral Responsibility

Oxford University Press, New York 2019, 174 pages

by Lorenzo Testa

In the philosophical debate on free will and moral responsibility, the analysis of agents acting in non-standard conditions has always been employed in order to reveal important features of our concept of agency. Those who are interested in free will and moral responsibility should be familiar with discussions about agents coerced in performing an action and agents who lack the possibility to do otherwise. In this 'little book', as he himself dubs it, A. Mele gives a precious contribution to the debate on moral responsibility and free will by analyzing a different kind of agents. These agents analyzed by Mele are those manipulated into performing some actions. An agent counts as manipulated when another agent – the manipulator – adjusts beforehand the conditions under which the manipulated agent will act, in such a way that the manipulated agent will perform a certain action or a series of actions. The analysis of such agents, as the author argues in the book, can shed light on the concept of moral responsibility. The reader, however, should not expect to find a defense of a full-blown theory of moral responsibility. The aim of the book is narrower, as it focuses on what manipulated agents reveal about the concept of moral responsibility. What they reveal is the relevance of agential history for ascriptions of moral responsibility. Accounts of moral responsibility that identify an historical component claim that the way an agent came to be in the internal condition on which he acts in a certain time is relevant to the ascription of moral responsibility for actions. These accounts accept a form of conditional externalism, the theory that claims that an agent may be responsible for A at least partly because of how he came to be in the internal condition that issues in his A-doing. Conditional externalism is the theory defended by Mele against conditional internalism. According to conditional internalism, if an agent finds himself in a certain condition C when he performs A, and he performs A because of a part P of C, then the agent is morally responsible for A no matter how he came to be in C. It is important to notice that the choice between internalism

ISSN (print) 2281-8618-ETS

and externalism does not mirror the choice between compatibilism and incompatibilism about determinism and moral responsibility. On the contrary, one of the advantages of Mele's defense of externalism is that his theoretical position should be adopted by both compatibilists and incompatibilists. In other words, forms of conditional externalism are to be preferred over forms of internalism, independently from the truth of compatibilism or incompatibilism. Thus, any plausible account of moral responsibility for actions should always include a reference to the history of agents who perform these actions. The author defends this thesis trough all the six chapters in which the book is divided. In the first chapter, Mele introduces the most relevant concepts that he will adopt in the rest of the book, e.g., determinism. The second and the third chapters include a series of thought experiments involving manipulated agents, as well as a reply to M. McKenna and M. Vargas on thesis defended by Mele in his previous works. In the fourth and in the fifth chapters, Mele reinforces his thesis, arguing that both compatibilists and incompatibilists should accept it. In the sixth and final chapter, Mele summarizes his argument and anticipates some objections. An interesting Appendix closes this work with some empirical experiments conducted by the author about non-specialists' intuitions about the thought experiments presented through the book.

I have mentioned that the defense of forms of conditional externalism is the upshot of the analysis of manipulated agents. This deserves further considerations. In the second chapter of the book, Mele introduces two scenarios in which two different agents perform the same morally wrong action A, namely murdering an innocent. Since throughout the book the author frequently refers to these two scenarios, I will offer a brief description of them. The agent in the first scenario is the cruel Chuck, who performs A under standard conditions: he is not forced into A-ing, nor A is the only possible course of action available to Chuck. Moreover, murdering an innocent fits perfectly into Chuck's evaluational system. Chuck has not always been so cruel, but he worked towards the formation of a cruel character. Given that, he does not feel remorse after his wrongdoing. The agent in the second scenario is Sally, a young woman who has always tried to act in a morally right way. Contrarily to Chuck, throughout her life Sally has built an evaluational system in which performing an action such as A is not even an option for her. Despite that, Sally, after having been manipulated by a team of nefarious neuroscientists, murders an innocent. The neuroscientists manipulated Sally's brain so that her new evaluational system is exactly the same as that of Chuck. Both Chuck and Sally performed a morally wrong action, namely murdering an innocent. Moreover, at the relevant time they share the same

system of values. Even so, Mele claims that while Chuck is morally responsible for A-ing, Sally is not morally responsible for murdering an innocent. What justifies, according to Mele, this asymmetry in the ascriptions of responsibility? He answers this question by claiming that our intuitions about Chuck and Sally reveal that their agential history is relevant for ascriptions of responsibility, since the only relevant difference between the two agents is their agential history prior to Sally's manipulation. If Mele is right about this, then conditional externalism is true, and a full analysis of moral responsibility should include a reference to the history of agents.

The methodology followed by Mele is quite common in analytical philosophy works on moral responsibility. Usually, authors propose their thesis, then build a thought experiment in order to elicit a certain intuition in the reader, and finally give and argument in favor of the intuition highlighted in the thought experiment. This book makes no exception to this general structure. On the contrary, the two thought experiments involving Chuck and Sally are widely discussed in the book. Moreover, Mele proposes many variations of the two initial scenarios. The reader who is not used to discussions about increasingly sophisticated versions of thought experiments may not find Mele's argumentative strategy so compelling. What is undoubtedly remarkable in the book, though, is the attention that the author gives to the importance and the role of thought experiments and intuitions that stem from them. This is a merit of the book, especially since the legitimacy and the role of thought experiments in analytical philosophy has recently faced skepticism. Mele explicitly treats his thought experiments and the intuitions about moral responsibility they elicit as a groundwork in order to introduce his arguments. In other words, even if it is true that his thought experiments serve as a reference in the whole work, the book contains a series of convincing independent arguments. Thus, the theoretical heavy lift is not solely done by thought experiments. Moreover, the Appendix contains a series of interesting empirical experiments conducted by Mele and a team of psychologists. In these empirical experiments, Chuck's and Sally's scenarios – and variations of them – were submitted to non-philosopher adult individuals. The conductors of the experiments then asked their subjects what their intuitions were about the ascriptions of responsibility. As a support to his own thesis, Mele notes that the majority of the individuals shares his intuitions about moral responsibility. Precisely, the majority shares his intuition that Chuck is morally responsible for his wrong action, while Sally is not morally responsible for hers.

There are at least two related remarks that are worth noting. One is about the legitimacy and the utility of thought experiments involving cases of ex-

treme manipulation. The other is the question on the value and the weight of our intuitions. Beginning from the former, it is important to note that Mele himself addresses the problem. In the last part of his book, in fact, he lists twelve possible questions about the thesis defended in the book, with the aim of anticipating criticism. It is an admirable thing to conclude a piece of philosophical work by replying to possible objections. Most of the replies are convincing enough and they serve a clarificatory purpose in case the reader has some doubts about the most relevant passages of the book. The ninth question, however, touches a crucial problem, and Mele's reply does not seem as convincing as in the other cases. Question 9 (p. 138) asks why we should care about arguments involving fictional agents so distant from the actual world. Consider the team of nefarious scientists who eradicate Sally's evaluational system and replace it with a new evil one, thus manipulating Sally in murdering an innocent. Given that this kind of extreme manipulation never actually happened in our world, why should we refer to this scenario? Mele replies by arguing that both R. Kane and D. Pereboom use these kinds of scenarios in defending their incompatibilist theses. Mele then adds that he had never considered replying to Pereboom and Kane that their thought experiments are invalid because they are not set in our actual world. This, however, is not a satisfying answer to the initial concern about the validity of thought experiments. After all, Kane's and Pereboom's arguments may suffer from the very same problem, namely making reference to hypothetical scenarios which are set under implausible conditions. If we want to use thought experiments in order to highlight some interesting characteristics about our concept of moral responsibility, why should we trust our intuitions about situations so different from our actual world? Mele, however, does not only refer to Kane and Pereboom in order to justify his own methodology. He adds that, as long as metaphysical or conceptual questions are at issue – and this is the case of both Kane's and Pereboom's books– it is a legitimate move to refer to situations that cannot happen in the actual world. And even if, in this book, Mele does not want to offer a complete account of moral responsibility or free will, he does want to offer a defense of externalism. Moreover, it could be argued that those who criticize the legitimacy of thought experiments need to provide an argument in order to defend their skepticism. Without a more detailed argument, claiming that thought experiments which involve far-fetched conditions are not justified would be no more than an intuition. But nowhere in the book Mele claims that intuitions should not be trusted and avoided at all costs; quite the contrary, intuitions about agents in the scenarios Mele builds are the starting point of philosophical reflection. This leads to my second critical remark: why should we reject the intuition

against the legitimacy of thought experiments involving implausible scenarios? Of course, Mele is right in pointing out the fact that intuitions alone are not sufficient in philosophy and a valid argument is to be provided if one wants to press the criticism against thought experiments. Still, it seems that the project of building such a valid argument is far from being doomed, and it should not be easily dismissed as a simple intuition – especially because of the relevance that the author gives to intuitions.

As a conclusion, it should be noted that these remarks could be directed to great part of the philosophical literature on topics like free will and moral responsibility. Mele is well aware of these potential problems, and he admirably mentions them. A more compelling treatment of these problems would be much appreciated, and we hope to read some further works by Mele on these issues.

Lorenzo Testa
lorenzo.testa01@universitadipavia.it
Università di Pavia