Φιλοσοφια
Volume 18, 2:2017

MANIPULATION ARGUMENT AND THE TRAP-INTUITION[1]

Zsolt Ziegler
Department of Philosophy and History of Science
Budapest University of Technology and Economics

I will challenge the manipulation argument, aiming to argue for the incompatibility of moral responsibility and determinism. By examining the intuition behind the first premise, stating that manipulated agents are not responsible, it will turn out that this statement can be traced to the manipulators themselves, who intentionally set up a plan against their subjects. The second premise, which states that there is no difference between determinism and manipulation concerning responsibility, will be argued to be false. In the deterministic worlds, actions are determined by blind causation. However, under the manipulation theory, agents are determined by the manipulator. I claim that the first premise is true, but the second premise is false.

INTRODUCTION

In this paper, I will will try to show that the so-called manipulation argument, which says that moral responsibility is indistinguishable between the manipulation theory and determinism, is untenable. By examining the content of the first premise of the argument, which maintains that manipulated agents are not responsible, it will become apparent that this judgment can be traced to the manipulators themselves. They want their agents to perform some acts through subtle manipulation methods and for the sake of which the manipulated persons perform the acts as if they were not responsible at all. According to the second premise of the manipulation argument, there is no significant difference between determinism and the manipulation theory with regard to responsibility. I want to argue that in deterministic worlds, actions are determined by natural blind causation. In the case of manipuled agents, the agents actions are determined by the manipulators' intentions. Simply put, a manipulator *intentionally got me to do "x,"* but in a deterministic system, the system itself *caused me to do "x."*

Cases of manipulation have been introduced in order to pose a problem for compatibilist theories. Compatibilism is the view that even if determinism is true, people can be morally responsible for their actions even though they cannot avoid doing them. Usually, classical compatibilism requires that the agent needs to be

able to do as he or she pleases. In these theories, compatibilism argues that the agent acts in accordance with the determination of one's own will, or he or she is capable of acting according to his or her available reasons. An agent thus needs, on the one hand, to recognize reasons and, on the other, to be able to translate these reasons into actions. He or she has the capacity for rational self-control and is moved guided by his or her reasons:

> If someone does something because he wants to do it, and if he has no reservations about that desire but is *wholeheartedly behind* it, then – so far as his moral responsibility for doing it is concerned—it really does not matter how he got that way. One further requirement must be added...: the person's desires and attitudes have to be relatively *well integrated into his general psychic condition.* (Frankfurt, as quoted by Fisher 2002, 27)

However, an agent's responsibility may be compromised, if another agent, a manipulator, constraints him or her with some degree to do a particular action. Agents of this kind could be *manipulated* and *covertly controlled* by other agents and, yet, given the classical compatibilist account, should be judged free and responsible since, it is claimed in cases of manipulation, the rational self-control mechanism remains intact. However, many philosophers (Mele 2006, Kane 996, Pereboom 2001, Fischer and Ravizza 1998, 194-202, 230-39, Haji and Cuypers 2004, and Russell 2002) argue that agents who are manipulated and covertly controlled, are obviously not free and responsible despite the fact that they may possess a general capacity for rational self-control of the kind that compatibilists have described.

Alfred Mele (2006, 171-72) has offered an example in which a person, Beth, who has a wonderful moral character, is brainwashed by another. According to this example, Beth's character was replaced with Chuck's truly evil one. With Chuck's evil character having been implanted into her, Beth woke up in the morning and, by having Chuck's values *well integrated into her general psychic condition,* she *wholeheartedly* committed a horrible crime. The fact that she is manipulated may trigger our intuitions to exempt her from responsibility even if the requirements of a compatibilist account are satisfied, that would make her blameworthy.

The manipulation argument against compatibilism satisfies all the compatibilist requirements for responsibility and at the same time is meant to elicit the intuition that in such cases agents cannot be held responsible. It wishes to emphasize that the deliberative mechanisms created by various deterministic processes are not different from manipulation. The argument can be set out as follows:

> I. If an agent is manipulated to perform an act, then she does not act freely and she is not responsible for the action. [1st premise]
>
> II. Concerning free action and moral responsibility, there is no significant difference between the agent's act as a result of being

> manipulated in a way and the way any normal human acquires his or her deliberative mechanism and values in a deterministic universe. [2nd premise]
>
> III. So determinism precludes free action and moral responsibility.[Conclusion] (Mele 2006, 189; Pereboom 2001, 11)

First, I am going to briefly summarize compatibilist replies against the manipulation argument. I will describe historical and nonhistorical compatibilism. I will not go to the details of these criticisms for my only purpose is to map out compatibilist views in order to locate my position against the manipulation argument. I will also argue that the first premise of the manipulation argument is true. Indeed, I admit that manipulated agents are not responsible. However, I will claim that there is a morally significant difference between the agent's act as a result of manipulation and as a result of determinism. Therefore, the second premise of the manipulation argument is faulty and the conclusion does not necessarily follow.

REPLY TO THE MANIPULATION ARGUMENT BY COMPATIBILISTS

Historical and Nonhistorical Compatibilism

Compatibilists are of two groups: The nonhistorical, who challenge the first premise, and the history-sensitive compatibilists, who deny the second premise.

History-sensitive or historical compatibilists think that all cases of manipulation are genuinely different from a normal, causally deterministic course of events and claim that morally responsible agency is an essentially historical notion. They contend that the second premise must be false.

> Two agents who are nonhistorical duplicates at a time might very well differ with respect to their status as free and morally responsible depending upon differences in their respective histories—that is, depending upon differences in their "historical properties." Hence, for the historical compatibilists, the concept of moral responsibility is historical in the same way that the property of being a sunburn or a genuine dollar bill is historical. (McKenna 2012, 154)

One prominent theory is offered by Martin Fischer and Mark Ravizza (1998 208, 210-11, 238). In order to solve the problem of manipulation, they suggest that an agent possesses his or her deliberative mechanism leading to the action only if the mechanism has the *right history* or *causal origins*. Once the deliberative mechanism is altered by another agent, using artificial methods, the agent loses her responsibility. Nonetheless, if the mechanism is created by normal deterministic causal processes, then the *ownership* of the agent over her action is established. In case of manipulation, the agent does not *own* the mechanism. But under determinism, the agent has rational self-control called *reason responsiveness* (Fischer 2006, 230).

The problematic part of Fischer and Ravizza's account on responsibility is the question *what makes a deliberative process original* in this narrow sense as opposed to manipulation. Fischer (2006, 240) is aware of this problem and writes:

> [T]he structure of our theory of moral responsibility is similar to the structure of "reliabilist" theories of knowledge. In these theories, ascertaining whether an individual has knowledge involves holding fixed the actual-sequence belief-producing mechanism and asking whether it is "reliable"—whether,for instance, it tracks truth (in Robert Nozick's terms). Indeed, since Nozick offers no general account of mechanism individuation (of belief-producing mechanisms).

From this it follows that simply by examining the history of the deliberative process we could distinguish cases of manipulation from cases of self-determination.

Others, nonhistorical compatibilists, however, argue that if manipulation is sufficiently complete and detailed, the manipulated agent acts of her own free will and must be morally responsible for what she does. This approach, thus, simply denies the first premise, by stating that manipulation does not undermine the agent's responsibility. Nonhistorical compatibilists hold that history is irrelevant but what counts for responsibility is whether the agent at the very moment of action has the relevant deliberative mechanism or not.

> Nonhistorical compatibilists are committed to the view that any two agents who are nonhistorical duplicates by virtue of sharing all of their "nonhistorical", "snapshot" or "current timeslice properties" do not differ with respect to their status as free and morally responsible. (Mckenna 2012, 154)

Frankfurt's (see Fisher 2002, 28) position is a clear-cut example for nonhistoricism:

> It is irrelevant whether those causes are operating by virtue of the natural forces that shape our environment or whether they operate through the deliberate manipulative designs of other human agents. We are the sorts of persons we are; and it is what we are, rather than the history of our development, that counts.

Nonhistorical compatibilism admits that manipulated agents are responsible. Robert Kane (1996, 67) has suggested that hard compatibilists are willing to "bite the bullet." It runs contrary to our common sense intuition that manipulation compromises an agent's responsibility. Therefore, hard compatibilists must deny that manipulation undermines moral responsibility. Accordingly, if it turns out that I am manipulated, I still have no reason to change my fundamental conception of myself as a responsible person.

Externalism and Internalism

The literature generally refers to historical compatibilism as externalism while nonhistorical compatibilism is often called internalism. Some use these terms interchangeably since history as the only external factor has been considered so far. However, these notions do not overlap entirely. Of course, the pair of externalism and internalism differs from the pair of historical and nonhistorical compatibilism in the following way. Externalist theories look to external or nonmental features of an agent's deliberative mechanism. Contexts under which agents come to perform their acts might differentiate their status of responsibility. This approach suggests that agents are responsible only under certain conditions. Some of these conditions, which are external to the agent, are responsibility-undermining while others are not. Externalism is also a compatibilist position. It tries to show that determinism is consistent with those responsibility-establishing circumstances combined with a one-way classical compatibilism, whereas manipulation described by the argument above is not. Nonetheless, externalism is a broader notion than historical compatibilism. In that broader sense externalism does not imply history-sensitive compatibilism. History is external to the agent but history is not the only possible external factor that can be relevant concerning responsibility. I will suggest a new one, a social factor.

Internalist theories put emphasis exclusively on the internal structure of an agent's mental properties that are supposed to account for morally responsible actions. Internalist compatibilist theories are also *current time-slice theories* like nonhistorical compatibilism. Similarly, this approach requires only the existence of a certain mental structure at the very time of actions. For an internalist compatibilist what preserves moral responsibility is nothing but a proper mental content. No matter how the content has been acquired, once one has it, he or she is responsible. However, internalism is a wider category than nonhistoricism. In fact, nonhistorical compatibilism is worth seeing as a subclass of internalism. Not surprisingly, possessing an internalist position does not necessarily imply nonhistorical compatibilism in a broader sense. Internalism concurs with almost every aspects of nonhistoricism except that it does not have to deny the relevance of history.

Internalist history-sensitive compatibilism

Exactly this little gap between internalism and historicism has been utilized by Manuel Vargas (2006) in creating a third alternative for an *internalist history-sensitive compatibilism*, which he calls *semi-structuralist compatibilism*. Vargas (2006, 364) accounts for a collection of capacities to foster the ***b**asic **a**gential **s**tructure of **r**esponsibility* (BASR):

> At least minimal rationality, sensitivity to justified moral norms, responsiveness to moral reasons, and the presence and normal operation of basic psychological features, including beliefs, pro-attitudes, and intentions, are surely some of the features of agency we are justified in fostering.

Vargas's positon is somewhere between internalism and historical compatibilism. On the one hand, it is supposed to avoid biting the bullet by putting emphasis on the importance of history. On the other hand, the right agential structure grounds moral responsibility. Once one has the right stucture, the considered agent turns out to be responsible. Internalist history-sensitive compatibilism says that the manipulated agent can be responsible so long as she retains the relevant structural faculties. According to Vargas (2006, 366-67):

> [I]f BASR is present in a Brave New World [case of manipulation], the agent ought to be counted as a responsible agent because she has the capacities we are justified in fostering through moral influence.

Whenever all the responsibility-ensuring conditions involving historical components are satisfied, compatibilists have no reason to deny the agent's responsibility. If the process of manipulation extends to historical features by which the manipulator creates the same kind of BASR in the same way as deterministic events usually do, then the agent is responsible. If the manipulation extends through time, it is no different from deterministic value acquisition. In this manner, an incompatibilist can satisfy the second premise of the manipulation argument. If so, Vargas cannot defeat the manipulation argument. However, Vargas, in the case of semistructuralism, denies the possibility of, say, responsibility retaining instant manipulation when an agent's entire value system is restructured in a moment. BASR, then, requires historical characteristics of internal structure but the structure of agency grounds responsibility. Thus, semistructuralism is an internalist history-sensitive account.

THE TRAP-INTUITION

In this section, I am going to examine the intuition suggesting why manipulated agents are not responsible. I claim that manipulated agents are not responsible because manipulators *intentionally* set up a plan against their subjects. I argue that we hold responsible the one bearing the intention of the act (Russell 2013). The manipulator wants an end, for the sake of which the manipulated person performs his or her act for which he or she is not responsible. I think in cases of manipulation we have a certain *trap-intuition* suggesting that the agent is not responsible. Furthermore, I understand manipulation as *doing A* in order for another person to perform *doing B*, or with the intention of making the agent do *B* without making him or her aware of the influence.

To exemplify the *trap-intuition,* let us suppose the following case. Peter desperately loves Klara who is the wife of Joseph. Peter is an excellent judge of character, so he knows all weaknesses of Joseph's character. Furthermore, Joseph has a weak will and he can be an easy subject of manipulation. Peter knows that Joseph always had difficulties with being faithful, although he has not cheated on Klara, yet. Peter also knows that Klara is a kind of person who cannot tolerate adultery. Hence, whenever Klara becomes aware of her being cheated, she will

immediately take the initiative to divorce Joseph. So, Peter sets up his plan to trap Joseph by creating a condition, say, "C," that would lead Joseph in a path that would cheat on Klara. "C" involves circumstances in which Joseph loses his ability to judge his own feelings on Klara properly. For example, in accordance with "C," Peter brings Joseph to a bar where Joseph meets a very attractive woman who flirted with him. He does not know, of course, that Peter secretly hired this woman. Peter tries to influence Joseph's way of thinking and step by step leading him to forget his feelings, at least temporarily, with Clara until ultimately he was placed in a situation, through Peter's inkling and the use of too much liquour, to forget himself and his better judgment and eventually cheat on Klara. Peter takes care that Klara will eventually know about her husband's cheating on her. To make the story short, Klara filed a divorce with Joseph and the two separated.

The intuition that Joseph is not morally responsible for cheating on Klara hinges on the fact that Joseph simply did not know what he was doing: the liqour and the flirting of the woman incapacitated his moral perception of proper values. This is what Peter wants to happen by his plan of manipulation. This manipulative situation meets Vargas's requirements (BASR) for responsible agency. Although BASR conditions are satisfied, there is an intuition here suggesting that Joseph is not responsible despite the fact that he retains a proper agential structure and the manipulation extends in time. However, I think that Peter's plan and his trap leading Joseph to cheat on Klara deprives Joseph of that degree of responsibility (Nelkin 2016).

Let us now suppose that in a different possible world, Joseph_2—the counterpart of Joseph—also cheats on his wife Klara_2. Joseph_2 goes through the events of "C" in the very same way as Joseph did except that the events of "C" happen by bare accident without the intervention of Peter. For example, Joseph_2 also meets a very attractive woman by accident in a bar who is flirting with him. Joseph_2's way of thinking and his capacity to evaluate his values and feelings towards his wife and towards the attractive woman happen to be distorted step by step and day by day until Joseph_2 cheats on Klara_2. Klara_2 figures out that Joseph_2 has cheated on her and so eventually they have a divorce. I think Joseph_2, in this case, is clearly responsible for cheating on Klara_2.

Behind the *trap-intuition* there are at least two reasons that suggest the manipulated agents are not responsible. The first is that the manipulator takes the advantages of the subject's weaknesses. The manipulator knows how best to manipulate the agent secretly, knows the factors by which the subject can be led to do certain things. The second reason is that the manipulated agent is trapped. Certain circumstances are *intentionally* arranged and set up in order to distort the agent's general capacity to evaluate his or her moral status in the case properly. However, in the second case, without Peter's trap, $\text{Joseph}_2$was unfaithful by his own volition. Joseph_2 cheated on her wife without any purposeful manipulation. Nonetheless, Joseph and Joseph_2 went on through the same circumstances "C" and experienced the same. Both Joseph and Joseph_2 control their actions in the very same compatibilist-like way. The only difference between them is the presence of Peter's *intentional* trap that deprives Joseph of direct responsibility.

FINAL CAUSATION IS RESPONSIBILITY DEPRIVING

There is a definite significant difference between the agent's act as a result of being manipulated and the way any normal human acquires his or her deliberative mechanism in a deterministic world. The difference between an agent being determined in a deterministic world because of *different notions of causation* are involved for describing cases of manipulation and determinism. To illustrate the difference between the two types of causes, I will now apply the Aristotelian notion of causes. Aristotle (*Phys*. 195 a 6-8. Cf. *Metaph*. 1013 b 6–9) distinguishes the efficient cause, i.e., "the primary source of the change or rest" from the final cause, i.e., "the end, that for the sake of which a thing is done."

Deterministic states of affairs are caused only by efficient causes. Nature—in the case of determinism—does not have any purpose or end that he or she intends to follow. It is blind and unpurposeful. However, cases of manipulation are caused by final causation which is governed by certain intentions aiming at a certain goal.

Importantly, these two notions of causes do not contradict each another. Once a person is manipulated by a manipulator, who follows an end for the sake of which the manipulation is done, such can also be described by an efficient cause in the sense of having a primary source of change. But, in this case, the manipulated agent is also caused to do certain things by final causation. Final causation, in the case of manipulation, is responsibility depriving because the manipulated person is intentionally governed by malevolence or bad benevolence. This character of being manipulated elicits the intuition that the agent is not responsible. He or she is trapped without even feeling the trap. There is no such intuition in the case of a deterministic agent. I admit, however, that from the frameworks in which the deterministic and manipulated agents may be viewed, there appears to be no empirical difference in the outcomes of their actions. Joseph_2 does not feel any determinism in his action but believes he performs in accordance with his own free will and volition. Joseph does not notice the manipulation but is brought into the act by the devaluation of his values and the slow transformation of his thinking from proper into the improper one.

This paper denies the second premise of the manipulation argument by maintaining that there is a significant difference in the actions of Joseph and Joseph_2 in the sense that—from the perspective of blameworthiness—we can directly blame Joseph_2 for his divorce with Klara_2. We cannot do the same with Joseph. Rather, we can directly blame Peter, the manipulator of Joseph, for the latter's action. In the case of the manipulation argument, the efficient cause originates from Peter and the final cause can be traced to Peter. In other words, in the case of manipulated Joseph, final causation is responsibility depriving.

Compatibilists attempt to save moral responsibility in harmony with efficient causation. Though no action is avoidable in the case of determinism, it is a matter of luck what traits and dispositions one may have and what circumstances in which one finds oneself that determine her moral development. Compatibilists (Fischer and Ravizza 1998, Wolf 1990) argue that agents can have control over their actions in the sense required for responsibility even though they do not have control over the causal factors of those actions. Note that blind efficient causation is the one

which is taken for granted in these compatibilist accounts. It is a matter of pure luck (without any purpose) what sort of traits one is born into and the circumstances in which he or she grows up. This is not true in manipulation when one's character traits are manipulated by purpose (see the "Zygote argument"[2] by Mele 2006) and when one's circumstances are set up for a certain end (see Derk Pereboom 2001, with his "Four-case argument for incompatibilism"[3]). The intuition presented in these cases seems to support the case of Joseph. Manipulated Joseph did not seem to be responsible while blindly determined $Joseph_2$,who thought he was free, did. Manipulated Joseph cheated on his wife apparently on conditions beyond his proper control, while $Joseph_2$, despite his deterministic background, apparently acted in full control of his wits.

Note that, this argument is an externalist and a nonhistory-sensitive argument against the manipulation argument. It is an externalist argument since it argues against the second premise but not in the way how history-sensitive compatibilism does.

CONCLUSION

In this paper, I examined the manipulation argument against compatibilism and various criticisms of it. I tried to explain the reason behind the first premise of the manipulation argument, stating that manipulated agents are not responsible in that they are captured in a trap situation. We intuit that manipulated agents contrary to their own intentions are trapped and deceived by a certain purpose. This cannot be true for determinism. Determinism is blind not having any purpose or goal. If ever $Joseph_2$, despite his past historical determinism, acted in a purposive way, it is because he thought he had his free will and volition. Hence, the rejection of the second premise renders the manipulation argument invalid.

NOTES

1. This research was funded by the Hungarian Scientific Research Fund OTKA K-109456. I am grateful to Tihamer Margitay, Istvan Danka, and Ákos Gyarmathy for their helpful comments on an earlier version of this paper.

2. The argument presupposes an example in which an entire agent is created "in utero," as Mele puts it (2006, 188). In a deterministic world, a divine person, Diana, created a zygote that was born and named "Ernie." Thirty years later, Ernie performed what Diana exactly intended Ernie to perform at the very time he did it. Suppose also that Bernie, who ws very similar to Ernie, but went through his life in a normal deterministic world, performed necessarily a similar act, like what Ernie did, at the relevant time. Mele poses the question whether a compatibilist can consistently say that Bernie is morally responsible when he acts while at the same time saying that Ernie is not?

3. Pereboom's example of manipulation involves four cases. In Case 1, an evil neuroscientist creates a humanoid with remote controls in its brain and causes it to kill a person. In Case 2, some neuroscientists also build a humanoid with a

computer for a brain and program it to be a murderer. In Case 3, a real human being is conditioned by rigorous behavior modifications to become a murderer. Finally, in Case 4, the murderer is a normal human being who grew up in a world where physical determinism is true. Pereboom suggests that the agents are not responsible in Cases 1-4.

REFERENCES

Aristotle. 1984. *Physics* and *Metaphysics*. In *The complete works of Aristotle.* Revised Oxford translation. Edited by Jonathan A. Barnes. Oxford: Matt-Pseudo. (Monograph Collection)

Fischer, John Martin. 2002. Frankfurt-style compatibilism. In *The contours of agency: Essays on themes from Harry Frankfurt*. Edited by Sarah Buss and Lee Overton. Massachusetts: MIT Press.

_____ . __. 2006. *My way: Essays on moral responsibility.* Vol. 57. Oxford University Press.

____ _ .__ and Mark Ravizza. 1998. *Responsibility and control: A theory of moral responsibility.* Vol. 61. Place: Cambridge University Press.

Haji, Ishtiyaque and Stefaan E. Cuypers. 2004. *Moral responsibility and the problem of manipulation reconsidered.* International Journal of Philosophical Studies 12 (4).

Kane, Robert H. 1996. *The significance of free will.* Vol. 110. New York: Oxford University Press.

McKenna, Michael. 2012. Moral responsibility, manipulation arguments, and history: Assessing the resilience of nonhistorical compatibilism. *Journal of Ethics* 16 (2).

Mele, Alfred R. 2006. *Free will and luck.* Vol. 10. New York: Oxford University Press.

Nelkin, Dana K. 2016. Difficulty and degrees of moral praiseworthiness and blameworthiness. *Nous* 50 (2).

Pereboom, Derek. 2001. *Living without free will*. Vol. 3. Cambridge: Cambridge University Press.

Russell, Paul. 2013. Selective hard compatibilism. In *Action, ethics and responsibility: Topics in contemporary philosophy*. Edited by Joseph Campbell, Michael O'Rourke, and Harry Silverstein. Vol. 7. Massachusetts: MIT Press.

Vargas, Manuel. 2006. On the importance of history for responsible agency. *Philosophical Studies* 127 (3).

_____. 2013. *Building better beings: A theory of moral responsibility*. Oxfrd: Oxford University Press.

Wolf, Susan. 1990. *Freedom within reason.* Oxford: Oxford University Press.

Submitted: 17 May 2016; revised: 23 March 2017