**Compatibilist Theories of Moral Responsibility and History: A Defence of Non-historical
Compatibilism**

By

Matthew O'Leary

A thesis

submitted to the Victoria University of Wellington

in fulfilment of the requirements for the degree of

Master of Arts

in Philosophy

Victoria University of Wellington

2015

Abstract

In the recent philosophical literature, a debate has emerged between compatibilists over the nature of moral responsibility. Though they agree that agents need to exercise control over their actions to be morally responsible for them, some compatibilists have argued that history is a necessary component of moral responsibility. That is, historical compatibilists believe that an agent needs to have the right kind of history leading up to an action for that action to authentically belong to an agent. In this thesis, I will address the important arguments for historical compatibilism and argue that the position is incorrect. The position I will defend holds that exercising control over your actions is necessary and sufficient for moral responsibility and that considerations of history are not necessary. The position I defend is referred to as 'non-historical compatibilism.'

Acknowledgements

I could not have completed this thesis without the support I have received. Special thanks to my supervisor Dr Ken Perszyk, who provided much insight and encouragement, as well as countless comments, critiques, and suggestions.

Thanks to the staff and the postgraduate community in the Philosophy department at Victoria University of Wellington. In particular, I would like to thank Josef Holden and Matt Macdonald for their thought-provoking discussions.

Financial support for this project was gratefully received from Victoria University of Wellington.

Finally, I would like to thank my family and friends for their unwavering support.

Contents

Chapter I: Introduction	6
1.1. Introduction	6
1.2. Definitions	9
1.3. Compatibilist Accounts of Control	11
1.4. Non-historical and Historical Compatibilism	13
1.5. Harry Frankfurt's Non-historical Compatibilism	16
1.6. Historical Compatibilism	19
1.6.1. Mele's Negative Historical Compatibilism	20
1.6.2. Fischer & Ravizza's Positive Historical Compatibilism	22
1.7. Conclusion	24
Chapter II: The Argument from Tracing	25
2.1. Introduction	25
2.2. The Argument	26
2.3. Why the Argument from Tracing Does Not Support a Necessary Historical Condition	28
2.4. Why the Argument from Tracing Does Not Support a Sufficient Historical Condition	35
2.5. Conclusion	38
Chapter III: Explaining Manipulation Arguments for Historical Compatibilism	40
3.1. Introduction	40
3.2. The Nature of Manipulation	41
3.3. Loaded Terminology	43
3.4. Global and Local Manipulation	45
3.5. Mele's Local Manipulation Case	47
3.6. The Structure of a Manipulation Argument for Historical Compatibilism	49
3.7. The Strength of Manipulation Arguments	51
3.7.1. Their Strength as Derived from the Transfer Principle	52
3.7.2. Mele's "Radical Reversal" Suggestion	54
3.7.3. Intuitions and Manipulation Arguments	58
3.8. Conclusion	60
Chapter IV: Criticising Manipulation Arguments for Historical Compatibilism	62
4.1. Introduction	62
4.2. Five Unsuccessful Attempts to Respond to Manipulation Arguments	63
4.2.1. Intentional and Nonintentional Compulsion	63
4.2.2. Manipulated Beth is a New Person	64
4.2.3. Chuck as Responsible for More than Beth	66
4.2.4. Manipulation as a Moral Violation	67
4.2.5. Harrison's Attempt to Transform Historical Compatibilists Into Non-historical Compatibilists	69
4.3. Manipulated Histories	73

4.4. Intuitions about Manipulated Agents Performing a Good Action	84
4.5. Manipulation and the Consequentialist Aspects of Praise and Blame	86
4.6. Fixating on Agential Properties	89
4.7. Responsibility-preserving Manipulation	93
4.8. Mele's 'First Action' Clause	98
4.9. Manipulation As Not Relevantly Different to Being Raised Naturally	102
4.10. Instant Agents	107
4.11. Instant Agents and Local Manipulation	110
Chapter V: Conclusion	119
5.1. Summary	119
5.2. Moderate Historical Compatibilism?	120
5.3. The Case of Robert Harris	122
Bibliography	127

Chapter I - Introduction

1.1. Introduction

When we praise and reward a friend for their generosity towards us, or blame and censure someone for their odious actions, we normally take their behaviour at face value. That is, though we may question their motives for these acts, or their appalling character, we do not usually inquire into the origins of their character or from where their motive may have arisen. It is interesting that this convention is a part of our blaming and praising practices because our judgements about a person's freedom and responsibility can be revised from knowing their history. This point is quite strikingly demonstrated in the abridged case of Robert Harris, a man who was on Death Row for the murder of John Mayeski and Michael Baker, two 16 year old boys:¹

Harris and his brother Daniel had been looking for a car they could use to rob a bank, when they noticed the boys' car. Harris threatened the boys with his gun and told them to drive him and Daniel to a canyon. Once there, Harris explained that they were going to take their car for the robbery, that they were to report the car stolen, and that some of the money from the bank robbery would go to them. The boys agreed, but once Mayeski had turned around, Harris shot him in the back. Harris chased Baker down a hill and shot him four times. When he returned to Daniel, he saw Mayeski was still alive and shot him directly in the head. Harris then began swinging his weapons in the air and started laughing.

¹ The information of Robert Harris comes from Gary Watson, "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme," in *Agency and Answerability: Selected Essays*, ed. Gary Watson (Oxford: Oxford University Press, 2004), 235-42.

After they had left the scene of the crime, Harris joked that they should pretend to be police officers and tell the parents that their sons had been killed. When Harris saw that his gun had flesh and blood stains on the barrel due to his point blank shot, he remarked “I really blew that guy’s brains out” and began laughing again. Harris was imprisoned, where he was despised by the other prisoners. His fellow inmates planned to celebrate the exact moment of his execution because they were so glad to be rid of him. The State Deputy Attorney General said of Harris that “If this isn’t the kind of defendant that justifies the death penalty, is there ever going to be one?”

The cruel acts Harris performed, coupled with his gleeful reaction to them, make us want to condemn Harris for what he has done. The description of Harris’ murders and how he felt about them seem to be all we need to determine that he is blameworthy for what he has done. But this reaction changes when we learn of Harris’ upbringing:

Harris was born prematurely after his alcoholic father returned home drunk and beat his wife and kicked her. His father frequently beat his children and was convicted twice for sexually molesting his daughters. Harris’ mother began to blame Robert for the hardships she faced. When Harris would go to touch his mother for comfort, she would push him away or kick him. His mother became an alcoholic and was arrested for bank robbery. Harris was teased at school because he had a learning disability and a speech problem. Eventually, Harris wound up at a federal youth detention center for stealing cars, where he was raped several times and twice slashed his wrists in suicide attempts.

This short description of only some of Harris’ childhood is enough to temper our outrage at his actions. He no longer seems to be a remorseless psychopath, but a victim of his circumstances. Focusing on the history of Harris, and discovering how he came to be the kind of person who would perform horrific acts appears to influence our assessment of the

moral responsibility of Harris' actions in the present. It seems that his history does have a bearing on his actions.

I will explore the issue of whether moral responsibility is a historical notion. That is, do elements from one's past factor into our considerations when we are determining whether that agent is free and responsible for the actions they perform now? I believe the arguments ultimately end in favour of answering this question with a qualified 'no.'

Provided that the agent has the capacities necessary to satisfy our best theory of control relevant for free will and moral responsibility, I do not believe that there is a further historical requirement that states that an agent must have come to satisfy the theory's conditions in the right way. Applied to Harris' case, if he had all the abilities and mental functions our theory of free will and moral responsibility demands at the time of the murder, then the depressing facts about his childhood would not change our verdict; we would know enough to determine that he is blameworthy.

In later chapters I will present and evaluate various arguments for and against my non-historical view. Throughout the remainder of Chapter I, I will explore the concepts foundational to the debate between the historical and non-historical side. In Section 1.2, I will define the traditionally significant positions of the free will and moral responsibility debate that are relevant to the compatibilist position I endorse. Section 1.3 will contrast compatibilism with incompatibilism's understanding of the control condition. Section 1.4 will explain the differences between historical and non-historical properties. Finally, Section 1.5 will give an example of a non-historical compatibilist's position, and historical compatibilists will receive the same treatment in Section 1.6.

1.2. Definitions

To be *morally responsible for an action*, as I will understand it, is for the action to belong to the agent in such a way that she deserves praise if the action she performed was morally good, and blame if the action was morally bad. Along with blameworthiness and praiseworthiness, moral responsibility also justifies a wide range of *reactive attitudes*. The reactive attitudes are emotional responses to other human beings and include feelings of indignation, guilt, resentment, and gratitude.² I will define *free will* as the “unique ability of persons to exercise control over their conduct in the fullest manner necessary for moral responsibility.”³ This definition will suffice for our purposes. We will be primarily interested in the control aspect of these concepts, so it is useful to have a definition that explicitly acknowledges that the notion of control is the same for free will and moral responsibility.⁴ It will also allow us to use the terms interchangeably.

It is generally accepted that there are at least two kinds of conditions an agent must satisfy for that agent to be morally responsible. The first is an *epistemic condition*. This condition relates to what kinds of beliefs about their actions an agent must hold in order to be morally responsible, and when ignorance can excuse an agent from praise or blame. The second element of moral responsibility is the *control condition*. This condition states that an agent must have a certain level of control over her actions to be morally responsible for what she does. The control condition of free will and moral responsibility explains why I can

² P.F. Strawson, "Freedom and Resentment," *Proceedings of the British Academy* 48 (1962).

³ Michael McKenna, "Compatibilism," ed. Edward N. Zalta, *The Stanford Encyclopedia of Philosophy* (<http://plato.stanford.edu/entries/compatibilism/2009>).

⁴ It should be noted that not all philosophers believe the same notion of control is shared by free will and moral responsibility. For example, semicompatibilists believe that moral responsibility and causal determinism are compatible, but make no claims about the compatibility of free will and determinism, thus leaving them open to the possibility that the two concepts utilise different notions of control. See John Martin Fischer and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility* (Cambridge: Cambridge University Press, 1998), 51.

sometimes be legitimately held responsible for the results of my actions, but not blamed for the occurrence of an earthquake. This essay will be concerned with whether these two conditions are together sufficient for a correct account of moral responsibility, or if our theory must also include a *historical condition* in their account. The historical condition lists a set of prerequisites that make reference to how an agent came to satisfy the control condition.

Determinism, on one interpretation, is the thesis that a complete description of the world at a time, together with the real laws of nature, causally necessitate the subsequent states of the world.⁵ If our universe is deterministic, then “at any instant there is exactly one physically possible future.”⁶ *Indeterminism* is the thesis that determinism is false.

With these definitions, the major positions in the traditional debate surrounding free will and moral responsibility can be described. *Incompatibilism* is the view that free will and moral responsibility are incompatible with the truth of determinism. If a world contains free and morally responsible agents, then an incompatibilist holds that it must be an indeterministic world. In contrast, *compatibilism* is the position that free will and moral responsibility are logically compatible with the truth of determinism. A great deal of attention has been given to whether incompatibilism or compatibilism is correct. I will not be addressing these arguments and will simply assume that compatibilism is true.⁷

Furthermore, I will also assume that the world we live in contains agents who are free and

⁵ Peter van Inwagen, “The Incompatibility of Free Will and Determinism,” *Philosophical Studies* 27, no. 3 (1975): 186.

⁶ *An Essay on Free Will* (Oxford: Clarendon Press, 1983), 3.

⁷ I will focus only on compatibilist theories, and not comment on whether incompatibilist theories of freedom and responsibility must also be historical, though I suspect that if compatibilists need historical conditions, then for similar reasons incompatibilists will need them. It is interesting to note that some libertarians see the need for historical conditions on their theories. While detailing agent-causal libertarianism, Clarke states that “if historical requirements concerning the acquisition and maintenance of one’s reason-states are needed, we may suppose them required here.” in Randolph Clarke, *Libertarian Accounts of Free Will* (Oxford: Oxford University Press, 2003), 134.

responsible for at least some of the actions they perform. These assumptions together necessitate the falsity of positions such as hard determinism and hard incompatibilism, which deny the existence of free will.⁸

1.3. Compatibilist Accounts of Control

Traditionally, compatibilists and incompatibilists have interpreted the control condition as having two different requirements; an alternative possibilities requirement (AP) and a sourcehood requirement. Compatibilists and incompatibilists who accept that AP explains why an agent is or is not morally responsible for their actions are known as *Leeway Theorists*.⁹ AP holds that an agent is morally responsible only if she had freedom to choose between relevant alternatives. That is, an agent is morally responsible for A only if she could have done not-A. As the historical condition bears more directly on the sourcehood condition, I will primarily be focused on that. However, I believe there are counterexamples that demonstrate the falsity of the leeway theorist's thesis, though I will not discuss them here.¹⁰

Source theorists, who accept the sourcehood requirement, are concerned with what conditions are necessary and sufficient for a person to be an appropriate source of agency, so that they can be deemed responsible for their actions.¹¹ Source incompatibilists believe

⁸ Hard determinism is the position that denies the existence of free will, by holding that free will is incompatible with determinism and that determinism is true. Hard incompatibilism holds that free will only exists in an indeterministic world where there are agents who can cause acts because they are substances and not because of their mental states, but that this is not one of those worlds. Hard incompatibilism is defended in Derk Pereboom, *Living without Free Will* (Cambridge: Cambridge University Press, 2001).

⁹ Neil Levy and Michael McKenna, "Recent Work on Free Will and Moral Responsibility," *Philosophy Compass* 4, no. 1 (2009): 102.

¹⁰ For the infamous counterexamples to AP, known as the 'Frankfurt cases', please see Harry G. Frankfurt, "Alternate Possibilities and Moral Responsibility," *The Journal of Philosophy* 66, no. 23 (1969).

¹¹ Levy and McKenna, "Recent Work on Free Will and Moral Responsibility," 102.

that an agent cannot be free or responsible for her behaviour if determinism is true because her actions would not be truly up to her. If the Big Bang causally necessitates a chain of events that culminate in Britta taking a Spanish class, then source incompatibilists will not deem Britta responsible for that action, as the source of her action lies ultimately in the events of the remote past. Due to their sourcehood worry, incompatibilists are often seen as attempting to explain how an agent is responsible for an action and how she is responsible for being the kind of agent that would perform that action.

In contrast, source compatibilists believe you can be determined and still be considered the source of your actions due to the high degree of control you can possess at the time of the action. To render their account more plausible, compatibilists have pointed out the rich set of powers and capacities an agent can exhibit while living in a deterministic world. These abilities, what I will refer to as *compatibilist agential structure*, include being able to recognise and respond to reasons, being motivated by and identifying with values that they have reflected on, resisting being moved by desires that conflict with these values, critical self-reflection, and regulating one's behaviour in light of moral considerations.

Compatibilists believe it is reasonable to praise and blame individuals with these abilities for what they do. These are source compatibilist powers as it is possible to possess the full suite of these abilities and be determined by factors beyond our control.

Compatibilists have tended to focus on defending their account of sourcehood as necessary and sufficient for moral responsibility. However, a new breed of compatibilists have emerged, who insist that the compatibilist account of sourcehood is necessary, but that agents must also satisfy a historical condition before they are responsible for their actions. They believe that it matters, not only that you fulfil their control condition, but how you came to fulfil their control condition. John Martin Fischer & Mark Ravizza, as well as

Alfred Mele, have offered compatibilist theories with a historical bent. To fully understand their theories, I will now explain what counts as a historical and ahistorical property.

1.4. Non-historical and Historical Compatibilism¹²

Historical phenomena are phenomena that depend on the past being a certain way. Two objects may look identical, that is they share all the same 'current time-slice' or 'snapshot' properties, but we may still classify them differently due to their historical properties. If two burns look identical and cause the same amount of discomfort to their victims, it does not make them both a sunburn. That is because a sunburn is a historical phenomenon; whether a burn is a sunburn depends crucially on whether it was caused by the sun. Similarly, you may have two five dollar notes that look identical, but only one can be used in a legitimate transaction if the other has been counterfeited. Being legal tender is another example of a historical property.

Non-historical phenomena are phenomena that depend only on their current time-slice or snapshot properties. A phenomenon is non-historical if a change in its past has no effect on the phenomenon. Being red, shiny, or round are all properties that are non-historical. We do not need to know how a red ball came to be that way in order to say that it has the property of redness. So long as the ball has the property at t_1 , we can say that it was red during t_1 .¹³

¹² 'Structuralism' has also been used to describe the non-historical position e.g. David Zimmerman, "Born Yesterday: Personal Autonomy for Agents without a Past," *Midwest Studies in Philosophy* 23, no. 1 (1999). Fischer & Ravizza use "current time-slice model" to refer to non-historical compatibilism in Fischer and Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, 198. 'Internalism' and 'Externalism' are used to refer to non-historical and historical compatibilism respectively in Alfred R. Mele, *Autonomous Agents: From Self-Control to Autonomy* (New York: Oxford University Press, 1995).

¹³ For a good discussion of historical and non-historical phenomena, see Fischer and Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, 170-206.

It is important that we distinguish between something being genuinely historical as opposed to it being apparently historical. This will help specify the distinct ways historical and non-historical compatibilists conceptualise moral responsibility. An example of something being superficially historical is your doctor questioning your travel history and using this data to differentiate your illness from others that share its symptoms. Although she appears to be using a past fact to influence her judgement of your current disease, it is not a genuinely historical fact. This is because the past fact is being used to guide your doctor epistemically to help her understand your current illness. The past fact is important only in that it gives us information about your current state of health. If the doctor was able to gain complete knowledge of your current time-slice properties, then those details would have been sufficient to deduce what illness you had and there would have been no need to query you about your past.

Contrast that with the two burns that are identical, though one is a sunburn. The fact that the burn was caused by the sun does not affect its surface area, the intensity of the pain, or any of its other snapshot properties. But whether it is a sunburn does depend on past facts about the burn's causal history; information that is not coded within the snapshot properties of the burn. The history of the burn does affect what kind of burn it is without impacting on its snapshot properties, unlike your illness, which in theory could have been determined if we knew enough about your snapshot properties. The past facts about the illness are only apparently historical because they only matter epistemically, whereas the facts about the burn are genuinely historical because they affect whether it is a sunburn.

Historical compatibilists hold that moral responsibility is essentially a genuinely historical concept. Past facts can alter whether someone is morally responsible without causing a change in their snapshot properties. Though two agents may be duplicates in terms of their

snapshot properties (they have the same compatibilist agential structure at a time), a historical compatibilist does not believe this is enough to ensure that both are morally responsible agents. It may be that one has the sort of history that the historical compatibilist believes nullifies moral responsibility. So for any two agents identical in their snapshot properties, it is open to the historical compatibilist to consider them differently, based on their history. As mentioned above, Fischer & Ravizza, and Mele are historical compatibilists.¹⁴

Non-historical compatibilists believe that historical differences between agents do not matter; moral responsibility is a non-historical phenomena. A non-historical compatibilist will take their favourite set of powers an agent can possess in a deterministic universe (such as the ones mentioned in Section 1.3). If the actor has the capacities, then they are morally responsible. If not, then they are not morally responsible. No discussion of how the actor came to possess these powers is necessary. Because of this, for any two agents that are identical in terms of their snapshot properties, the non-historical compatibilist must give equivalent judgements concerning their status as morally responsible agents. Harry Frankfurt is the most prominent non-historical compatibilist, whilst Michael McKenna is the most sympathetic to the position in recent literature.

Though there is an intramural debate over whether moral responsibility is genuinely historical, both types of compatibilists can accept that apparently historical facts can determine moral responsibility. Facts about the past can provide us with information about whether you are in control of your actions. For example, if I know you have been given a hallucinogenic drug this morning, then this past fact tells me that you are not connected to

¹⁴ It should be noted that Mele is agnostic about the truth of compatibilism and is only putting forward what he believes to be the best version of the theory. For the purpose of this thesis, he will be treated as having endorsed historical compatibilism.

reality in the right way to be praiseworthy or blameworthy for your actions. Both types of compatibilists believe apparently historical facts can inform us as to whether an agent is responsible or not, but only historical compatibilists utilise genuinely historical facts.

It is important that I correct a potential misunderstanding of the debate between non-historical and historical compatibilists. The words 'current time-slice' and 'snapshot' are often used to describe the properties a non-historical compatibilist believes constitute moral responsibility. I will also adopt this usage, but the wording suggests that non-historical compatibilists accept that all the properties relevant to being morally responsible are instantaneous. If this were true, then all historical compatibilists would need to do to refute the position is demonstrate that value formation, or deliberation or reflecting on your desires are all events that take time to happen. But this is not how non-historical compatibilists construe freedom and moral responsibility. They understand that those events have temporal extension. The debate between non-historical compatibilism and historical compatibilism is whether the properties that satisfy the control condition for actions are sufficient for free will and moral responsibility, or whether facts about how the agent came to satisfy the control condition are necessary.¹⁵

1.5. Harry Frankfurt's Non-historical Compatibilism

Harry Frankfurt puts forward a theory of moral responsibility that falls within the category of *mesh theory*. Mesh theories posit that for an agent to be morally responsible for actions,

¹⁵ See also Gary Watson, "Some Worries About Semi-Compatibilism: Remarks on John Fischer's the Metaphysics of Free Will," *Journal of Social Philosophy* 29, no. 2 (1998): 139-41; "Soft Libertarianism and Hard Compatibilism," *The Journal of Ethics* 3, no. 4 (1999): 362-64; David Zimmerman, "That Was Then, This Is Now: Personal History Vs. Psychological Structure in Compatibilist Theories of Autonomous Agency," *Noûs* 37, no. 4 (2003): 641-42; "Born Yesterday: Personal Autonomy for Agents without a Past," 240-41; Manuel Vargas, "On the Importance of History for Responsible Agency," *Philosophical Studies* 127, no. 3 (2006): 355-56.

there must be the right kind of harmony between some of the agent's mental states and the actions they produce.¹⁶ Though I will not explicate the specific details of Frankfurt's theory, it is a kind of *deep-self view*.¹⁷ The deep-self view holds that for an agent to be responsible for her actions, it is not enough that she simply performs the actions. Her actions must be an expression of her character, who she truly is deep down. Frankfurt's acceptance of mesh theory and the deep-self view means that he believes the mental states that must be in harmony with your actions are the ones with which the agent identifies. Our actions should express the characters and personalities we endorse.

For example, consider an unwilling drug addict. The drug addict wants to stop desiring the drug, or values a life free of addictive substances. However, her body continues to produce irresistible cravings and though she tries to resist, she relents and takes the drug again. Frankfurt's view holds that this person is neither free nor responsible for her drug taking, and so it is not reflective of who she really is; there is not the right mesh between her values and actions. Had her deep-self been in control of the action, she would have elected to not take the drug.¹⁸

Frankfurt's position is non-historical, so it only refers to the mental economy of the agent at the time of action and is not concerned with how the agent came to possess the deep-self she ultimately ends up having. Recognising this, Don Locke developed a counterexample to Frankfurt. Suppose a devil or neuroscientist deletes your deep-self and engineers a new one within you. Frankfurt would still deem you morally responsible for any acts resulting from this newly acquired deep-self because all he requires is that there is the right relation

¹⁶ Levy and McKenna, "Recent Work on Free Will and Moral Responsibility," 102.

¹⁷ For Frankfurt's account, see Harry G. Frankfurt, "Freedom of the Will and the Concept of a Person," *The Journal of Philosophy* 68, no. 1 (1971).

¹⁸ For a lengthier description of the deep-self view, see Susan Wolf, "Sanity and the Metaphysics of Responsibility," in *Responsibility, Character and the Emotions*, ed. Ferdinand Schoemann (Cambridge: Cambridge University Press, 1987).

between a deep-self and action. But such manipulated agents are intuitively not morally responsible. If I was a lifelong smoker with no intention to quit, and the devil altered my deep-self such that I wanted to quit, it seems that it would be odd to praise me if I subsequently went cold turkey and you knew these details.¹⁹

In reply to this case, Frankfurt controversially wrote:

[T]o the extent that a person identifies himself with the springs of his actions, he takes responsibility for those actions and acquires moral responsibility for them; moreover, the question of how the actions and his identifications with their springs are caused is irrelevant to the questions of whether he performs the actions freely or is morally responsible for performing them.²⁰

Moreover, as time went on and historical compatibilist positions were developed, Frankfurt continued to stick to his guns:

[I]t seems to me that 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. [...] As long as they are unequivocally attributable to him as his desires and his attitudes, it makes no difference – so far as evaluating his moral responsibility is concerned – how he came to have them.²¹

Frankfurt's non-historical compatibilist reply here has been seen not just as biting the bullet, but biting a mortar shell.²² His response does strike one initially as being highly counterintuitive and he does not provide much in the way of argument for these claims. Furthermore, McKenna, who has provided some defence of non-historical compatibilism, officially remains agnostic on the historical debate because he finds examples similar to Locke's persuasive.²³ In subsequent chapters, I will develop the points these theorists have

¹⁹ Don Locke, "Three Concepts of Free Action I," *Proceedings of the Aristotelian Society* 49(1974): 104-06.

²⁰ Harry G. Frankfurt, "Three Concepts of Free Action II" (1975): 121-22.

²¹ "Reply to John Martin Fischer," in *Contours of Agency: Essays on Themes from Harry Frankfurt*, ed. Sarah Buss and Lee Overton (Cambridge, Mass.: MIT Press, 2002), 27.

²² Michael McKenna, "Responsibility and Globally Manipulated Agents," *Philosophical Topics* 32 (2004): 171.

²³ "Defending Nonhistorical Compatibilism: A Reply to Haji and Cuypers," *Philosophical Issues* 22, no. 1 (2012): 277.

made and go beyond them by providing arguments for why non-historical compatibilism is true and why Frankfurt's claims are not as bitter a bullet to swallow as one might think. But alternatives to the non-historical compatibilist option will be presented first.

1.6. Historical Compatibilism

Historical compatibilism consists of two elements. The first is a compatibilist account of an agent's abilities that satisfies the control condition of free will and moral responsibility. This account could be filled out with reference to a deep-self, the ability to regulate your behaviour with moral considerations, etc. But in addition to the control condition, historical compatibilists believe that moral responsibility requires a historical condition; it is crucially important to a historical compatibilist how one came to fulfil the criteria of the control condition. Historical conditions come in two varieties; positive and negative. A positive historical condition is when an agent *must* have a certain kind of history in order to be considered morally responsible. I will call theories with this kind of condition *positive historical compatibilism*. As we will see, Fischer & Ravizza's historical compatibilism falls under this term. A negative historical condition is when an agent *must not* have a certain kind of history in order to be considered morally responsible. I will refer to these theories as *negative historical compatibilism*. Mele's historical compatibilism is of this stripe.²⁴

My main concern will be with historical compatibilist's historical condition, not with their choice of control condition. Furthermore, most of my arguments against historical compatibilism should apply, regardless of the specific details of their historical condition.

²⁴ Ishtiyaque Haji & Stefaan E. Cuypers also have a negative historical compatibilist account in the literature, though for our purposes, Mele's account will suffice. For a statement of their view, see Ishtiyaque Haji and Stefaan E. Cuypers, "Magical Agents, Global Induction, and the Internalism/Externalism Debate," *Australasian Journal of Philosophy* 85, no. 3 (2007).

What is important for understanding my arguments is the distinction between positive and negative historical conditions. In Chapter II, Section 2.4, I will briefly criticise compatibilists who believe that history is sometimes *sufficient* for moral responsibility. And in Chapter V, Section 5.2, I will argue against compatibilists who believe history can *mitigate* moral responsibility without eliminating it. However, my primary aim is to object to compatibilists who believe that some form of historical condition is *necessary* for moral responsibility.

1.6.1. *Mele's Negative Historical Compatibilism*

Before I can state Mele's negative historical compatibilism, it will be helpful to unpack a few terms. A *pro-attitude* is one's values, principles, preferences and desires.²⁵ Agents have a variety of mental capacities which means that they can reflect on and critically consider a pro-attitude before they adopt it. In contrast, a pro-attitude has *bypassed* your compatibilist agential structure when you have not accepted the pro-attitude because of your mental abilities, but have acquired the pro-attitude in spite of them.²⁶ A pro-attitude is *practically unshedddable* for an agent if changing or attenuating the pro-attitude is not a genuine option for the agent, barring certain extreme scenarios outside of the agent's control. Skyler's valuing her children's welfare is practically unshedddable if the only way she would be able to rid herself of the pro-attitude is if her children's lives were threatened and the only way to save them was to give up the pro-attitude.²⁷

I can now give Mele's negative historical condition on moral responsibility which states when an agent is neither free nor responsible (NFM).

²⁵ Mele, *Autonomous Agents: From Self-Control to Autonomy*, 151.

²⁶ *Ibid.*, 166-67.

²⁷ *Ibid.*, 153.

NFM. An agent who performs an overt action *A* does not freely *A* and is not morally responsible for *A*-ing, if

- (1) She expresses unsheddable values in *A*-ing;
- (2) Owing directly to those values, she could not have done otherwise than *A* in the circumstances (on a compatibilist reading of “could have done otherwise”);
- (3) Those values were very recently produced in a way that bypassed her capacities for control over her mental life by value engineering to which she did not consent and are seriously at odds with autonomously acquired values of hers that were erased in the process;
- (4) She retains no pre-existing value that is promoted by her having unsheddable values she expresses in *A*-ing; and
- (5) *A* is the first overt action she performs on the basis of his new values.²⁸

If your history includes *NFM*-style compulsion, then Mele claims your values are inauthentic and that you are not morally responsible when acting from that value. Condition (3) will be vital to an argument in Chapter 3, Section 3.7.2, as well as an argument in Chapter IV, Section 4.7. For now, it is significant to note that the implication of (3) is that if Locke’s devil from Section 1.5 were to abolish your old values and bestow new ones on you, you would still be morally responsible for your action if the new values were not “seriously at odds” with the values you had obtained on your own.

²⁸ *Free Will and Luck* (New York: Oxford University Press, 2006), 170; “Manipulation, Compatibilism, and Moral Responsibility,” *The Journal of Ethics* 12, no. 3-4 (2008): 268-69; “Moral Responsibility and History Revisited,” *Ethical Theory and Moral Practice* 12, no. 5 (2009): 470.

1.6.2. Fischer & Ravizza's Positive Historical Compatibilism

To understand Fischer & Ravizza's historical compatibilist account, I must first introduce the concept of a *mechanism*. A mechanism is simply the process which leads to the action, or the way an action is brought about.²⁹ For example, when Lucille selfishly blames her son for crashing the car when she was driving, that behaviour is issuing from her mechanism of practical reason. But when she chooses to drink alcohol, a different kind of mechanism is operative; her addiction to the substance. Examples of different kinds of mechanisms are practical reasoning, unreflective habit, hypnosis, and direct electrical stimulation of the brain.³⁰ Fischer & Ravizza do not attempt to individuate mechanisms, and presuppose that there is a way of identifying the relevant mechanism for a particular action.³¹ However, they do state that your central nervous system is a mechanism of a physical kind when it produces an irresistible desire for you to take a drug.³² This suggests that they believe mechanisms can be differentiated and specified based, in part, on how the process leading up to the action is physically instantiated.

To possess the control over your actions necessary for moral responsibility, Fischer & Ravizza state that your mechanism must issue in behaviour that is *reasons-responsive*. Very roughly, being reasons-responsive involves not only recognising and acting on reasons as they actually are in the current situation, but being able to recognise and react to different reasons in counterfactual circumstances. Suppose Sarah's mechanism of practical reasoning leads to her cheating on her partner with an old girlfriend because she correctly believes that it is unlikely she will be caught. Sarah is only reasons-responsive if, had it been the case

²⁹ Fischer and Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, 38-39.

³⁰ *Ibid.*, 233.

³¹ *Ibid.*, 40, 46-47.

³² *Ibid.*, 48.

that it was highly probable her spouse would have discovered the affair, she would have recognised that fact and chose not to cheat for that reason.

In addition to being reasons-responsive, Fischer & Ravizza believe an agent needs to satisfy a historical condition before they are responsible for what they do. More specifically, they believe that for an agent to be morally responsible for an action, the agent must *take responsibility* for that particular kind of mechanism, so that we may deem the mechanism as authentically belonging to the agent.³³ There are three requirements for an agent to take responsibility for a mechanism. Firstly, the individual must see that her actions and motivations have causal consequences on the world. Secondly, the individual must accept that she will be the subject of her community's reactive attitudes in certain contexts and that this is fair. She must realise that it is appropriate, at times, for others to express indignation, pride, etc., in response to some of the actions she performs. Finally, these beliefs the agent has about herself must be appropriately linked to the evidence she has for them.³⁴ Fischer & Ravizza's theory is a variety of *historical* compatibilism because they believe that even if an agent has the right kind of control over her action (in their terms, is reasons-responsive), she would not be morally responsible for the actions she performs unless she also had a history in which she had taken responsibility for her actions.

Any agent who does not satisfy their three conditions of taking responsibility is, by Fischer & Ravizza's lights, not morally responsible. Fischer & Ravizza do hold that, by and large, most ordinary agents have taken responsibility for their ordinary mechanisms, such as practical reasoning and so we can be legitimately held accountable for our actions, even if they are determined.

³³ Ibid., 210.

³⁴ This account is developed in *ibid.*, 207-39.

1.7. Conclusion

This chapter has sought to provide some background on the in-house debate between compatibilists as to whether moral responsibility is an essentially historical notion or not. The non-historical and historical compatibilist distinction has been presented, and I have elaborated the foremost accounts that have been developed in the literature. In the remaining chapters, I will analyse and critique the two major arguments for historical compatibilism. Chapter II is devoted to an argument that I refer to as ‘The Argument from Tracing,’ an argument that attempts to establish that our praising and blaming practices already support the notion that freedom and responsibility are historical. Chapters III and IV, the majority of my thesis, focus on the manipulation argument, which I believe to be the main pillar of support for historical compatibilism. Manipulation arguments are similar to Locke’s argument against Frankfurt that I discussed in Section 1.5. Historical compatibilists hold that non-historical compatibilists cannot reject the possibility that a manipulated agent could be free and believe we should not accept non-historical compatibilism for that reason. Chapter III will clearly explain the concepts and different versions of manipulation arguments that can be employed in this debate. Chapter IV will contain a multitude of arguments for why we should not accept the conclusion of the manipulation argument, but instead believe that manipulated agents are free and responsible for what they do.

Chapter II – The Argument from Tracing³⁵

2.1. Introduction

Typically, two arguments are given to support the notion that moral responsibility is a historical concept. One of these arguments, the manipulation argument, will be evaluated in Chapters III and IV. This chapter will focus on the other argument in the historical compatibilist's arsenal; what I will call the Argument from Tracing. Though this argument is in the literature, it is often quickly assumed to succeed, so it has not faced much in terms of in-depth scrutiny. I will present new objections to the historical compatibilist's argument and claim that it does not achieve what it sets out to accomplish. The Argument from Tracing begins with a case in which both non-historical and historical compatibilists would agree that the agent is morally responsible for the action. However, the best explanation for why the agent is responsible appears to involve some fact of their past. Thus the argument demonstrates that our pre-theoretical commitments support the view that history is relevant to moral responsibility.

Historical compatibilists disagree between themselves over what kind of historical condition the argument supports. Some appear to believe that it verifies a historical *necessary* condition, where even if you exercise control over the action, you must have the right kind of history to be deemed morally responsible.³⁶ Others believe it demonstrates that historical facts are sometimes *sufficient* for moral responsibility.³⁷ The intuitions in the

³⁵ This chapter has greatly improved thanks to comments from Alfred R. Mele.

³⁶ Fischer and Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, 194-96, 201-02, 07.

³⁷ Alfred R. Mele, "Moral Responsibility and Agents' Histories," *Philosophical Studies* 142, no. 2 (2009): 162-63; Manuel Vargas, "On the Importance of History for Responsible Agency," *Philosophical Studies* 127, no. 3

initial case of the argument are so strong that even those sympathetic to non-historical compatibilism, such as McKenna, have stated that any compatibilist theory must accept this sufficiency condition for it to be a credible theory of moral responsibility.³⁸

Despite its initial plausibility (and perhaps because of it), the argument has been largely neglected by non-historical compatibilist critics. In this chapter, I will argue that the Argument from Tracing fails to support either a necessary condition, or a sufficiency condition. Section 2.2, will contain an explanation of the Argument from Tracing. In Section 2.3, I will argue that if the argument establishes the historical compatibilist's necessary condition, it does so only if it also concludes that no human being is ever morally responsible. This is anathema to historical compatibilists, as they tend to affirm the belief that we are in fact sometimes morally responsible.³⁹ In Section 2.4, I will argue that though agents in the tracing examples are morally responsible, the argument does not support a sufficiency condition, as it is a mistake to view the tracing examples as employing genuinely historical facts. Throughout this chapter, I will use the deep-self view as an exemplar of a compatibilist theory of control.

2.2. The Argument

The Argument from Tracing rests on scenarios that are called 'tracing examples.' Imagine that two individuals, Beadie and Kima, are driving drunk to their respective homes. Both are incapacitated by their intoxication in the sense that their deep-selves are not governing

(2006): 356-57; *Building Better Beings: A Theory of Moral Responsibility* (Oxford: Oxford University Press, 2013), 273-75.

³⁸ Michael McKenna, "Moral Responsibility, Manipulation Arguments, and History: Assessing the Resilience of Nonhistorical Compatibilism," *The Journal of Ethics* 16, no. 2 (2012): 156.

³⁹ Mele, *Autonomous Agents: From Self-Control to Autonomy*, 254.

their behaviour. Because they are very drunk, Beadie and Kima both lose control of their vehicle and each runs into a toddler, killing both toddlers instantly.

Despite not fulfilling the control condition of moral responsibility at the time of the accident, it is not necessarily the case that this fact will excuse either agent of culpability. And though both committed the same act due to the same factors, it is possible that only one of the agents will be held responsible for what they have done. We are able to determine that Kima is responsible for the crash while Beadie is not by examining the events that led up to the crash.

Imagine that Kima was at a party. Though she knew that if she got drunk, there was a risk that she might drive, Kima still freely chose to consume alcohol. Unfortunately, Kima did drive, and killed a toddler. Compare this with Beadie, who was abducted and against her will, had alcohol injected into her bloodstream. She was then put into her car, made to drive home, and subsequently crashed into a toddler.⁴⁰

Even though they both perform the same action, only Kima is morally responsible for the death of the toddler. Kima's deep-self did not control her behaviour at the time of the crash, but she is still blameworthy for the action because we can trace her actions back to an earlier time, where she was sober and in control of her decisions. Before Kima began to drink any alcohol, she could reasonably have foreseen the possibility that she might drink and drive, and could control whether she drank or not. Therefore, she is responsible for the crash. By contrast, Beadie did not voluntarily drink and drive. We cannot trace back from the crash to a time where she was in control, could foresee the consequences of her act, and could freely choose to imbibe alcohol. These facts explain why we do not think Beadie is morally responsible.

⁴⁰ This example originally comes from Fischer and Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, 195.

All this appears to be grist for the historical compatibilist's mill. A historical compatibilist believes that moral responsibility is a historical notion and this is a case where everyone believes that Kima is responsible whereas Beadie is not, and this difference in judgement is not explained by the snapshot properties of Kima and Beadie at the time of the action. Beadie and Kima could be identical in all the ways that matter for the non-historical compatibilist's notion of moral responsibility. They could be equally intoxicated, equally able to regulate their behaviour with reasons when sober, have the same reaction time, etc. But still we would regard only Kima as morally responsible. And this appears to be because what is driving our asymmetrical responses to these cases is not the properties of the individuals at the time of the action, but their histories leading up to the action. Kima's history is responsibility-preserving whilst Beadie's is responsibility-undermining.⁴¹

2.3. Why the Argument from Tracing Does Not Support a Necessary Historical Condition

Fischer & Ravizza believe that the tracing cases from Section 2.2 can be used to defend a necessary historical condition.⁴² That is, they believe that the lesson of the tracing stories is that you must examine the history of an agent to determine whether she is responsible. They find the intuition from the tracing cases to be highly plausible, so they move quickly from the cases to the conclusion that moral responsibility is necessarily historical.⁴³

⁴¹ There is a concern that Fischer & Ravizza's examples may prime us towards thinking the Argument from Tracing is stronger than it actually is, because their example of a responsibility-undermining history is one that involves manipulation, which features prominently in the second main argument for historical compatibilism. While this may lead us to give undue weight to the argument, I do not believe this will affect my critique of it. Thanks to Ken Perszyk for pointing this out to me.

⁴² Fischer and Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, 207.

⁴³ *Ibid.*, 195-96.

However, I will question the argument's ability to support this conclusion. I will reconstruct the Argument from Tracing fully and in standard form to make clear the structure of the argument and how I wish to attack it. As it does not appear in standard form in Fischer & Ravizza's writing, I will pay extra attention to remaining faithful to their text.

The tracing examples demonstrate the following premise in the historical compatibilist's argument:

- (1) For any agent *S*, *S* is morally responsible for an action *A* at *t* only if *A* is linked appropriately to *S*'s compatibilist agential structure *CAS* at *t*.

(1) captures the intuition engendered by the tracing cases. Kima is morally responsible for crashing into the toddler because that act was indirectly caused by her CAS. (1), as currently stated, is a premise that enjoys approval from most compatibilists, including myself as it is one way of stating the compatibilist's control condition. Some historical compatibilists appear to interpret (1) as saying that moral responsibility for actions can be historical, in the sense that whether an agent's action is linked appropriately to her CAS can depend on past facts. I will accept this historical gloss of (1) presently, but will challenge it in Section 2.4.

The argument is intended to establish the truth of historical compatibilism, so its conclusion is just a restatement of the historical compatibilist's thesis:

- (C) Moral responsibility is essentially historical. That is, *S* is only morally responsible for *A* at *t* when *A* is linked appropriately to the agent's *CAS* at *t* (i.e. when they satisfy the control condition), and when that *CAS* has been brought about in the appropriate way (i.e. when they satisfy the historical condition).

But how does (C) follow from (1)? (1) states that when an agent is morally responsible for an *action*, that action must have an appropriate link to the agent's deep-self and that this link can extend over time. But (C) is concerned, not just with responsibility for actions, but with the nature of the *agential structure*, namely that it is essentially historical. Actions and capacities are vastly different phenomena, so even if historical compatibilists are correct in claiming that responsibility for actions must be historical, this fact is insufficient to demonstrate a claim about the relationship between moral responsibility and the nature of our deep-selves. We can accept that Kima is responsible for her act of drunk driving because we can trace it back to her deep-self, but this does not force us to accept a further historical condition concerning Kima's deep-self. The Argument from Tracing involves a logical leap until we are given a reason to believe that what applies to actions should also apply to CAS.⁴⁴

What is needed to make the Argument from Tracing valid is a premise linking (1) and (C). As this version of the Argument from Tracing has not been formally stated by historical compatibilists, it is unclear what they would have in mind for the task. However, from their writings, it appears they believe that (1) provides direct support for their historical condition (as opposed to providing mediated support, where (1) supports an additional premise that supports the historical condition).⁴⁵ Given that historical compatibilists move quickly from

⁴⁴ McKenna appears to endorse a reply of this kind when he says "[the Argument from Tracing], however, does not give us reason to think that the exercising of those very agential capacities constitutive of control require *further* historical constraints." See Michael McKenna, "The Relationship between Autonomous and Morally Responsible Agency," in *Personal Autonomy: New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy*, ed. James Stacey Taylor (Cambridge: Cambridge University Press, 2005), 219.

⁴⁵ I think this is apparent when Fischer and Ravizza write that the "tracing aspect of our theory of moral responsibility implies that the theory is genuinely historical" in Fischer and Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, 196.

(1) to (C), it appears they have implicitly been adopting the following premise to make the argument valid:

(2) If *S* is morally responsible for *A* at *t* only if *A* is linked appropriately to the agent's CAS at *t*, then *S* is only morally responsible at *t* when their CAS at *t* has been brought about in the appropriate way.

As the consequent of (2) is one of the conditions of historical compatibilism, it is able to link the ideas contained within the tracing examples to historical compatibilism and make the Argument from Tracing valid. But several problems loom large for the historical compatibilist who endorses an Argument from Tracing that includes (2). The first is that it appears that there is no independent motivation to accept (2), apart from the motivation to make the Argument from Tracing valid. The more serious problem I will focus on stems from the fact that (2) is supposed to enshrine the belief that facts about responsibility for actions in the tracing cases have a bearing on facts about CAS.

As the tracing cases indicate, we are responsible for our actions when they can be traced back to our deep-selves. And if we agree with (2), then we believe that this fact about an action's history shows us that the compatibilist control condition must include a historical element. But if (2) is the reason we accept a historical condition, then what kind of history is required for CAS? That is, what shape will the historical condition take, given that it is the tracing cases that ground it? We are only able to say that our deep-selves are historical in the same way that responsibility for actions is historical. Just as our actions must be linked back to our deep-self, our deep-self must itself be linked back to a previous deep-self.

There are two problems with accepting a historical condition of this stripe. Firstly, it is not clear in what way the present deep-self should be linked back to past deep-selves, especially if the connection is to mirror the tracing employed in the tracing examples. Perhaps a historical compatibilist will say that the previous deep-self must allow or endorse the values from which the current deep-self acts. Regardless of how the details are specified, it seems that the Argument from Tracing will motivate a historical condition quite different to the ones currently defended by historical compatibilists, such as the view that an agent must have taken responsibility, or that the agent cannot have any compelled values or desires.⁴⁶ The Argument from Tracing is not a good defence of the necessary condition because it is not obvious that it will establish a historical condition that has been argued for in the literature.⁴⁷

Even if that problem can be solved, historical compatibilists will be unable to affirm the Argument from Tracing's historical requirement without also being forced to hold that no human has ever been morally responsible for anything that they have done. This is due to the fact that this historical condition requires a deep-self to be linked back to earlier deep-selves. Kima's deep-self at the time of drinking must be connected back to an earlier deep-self for her to be deemed morally responsible. But then we can question the status of this earlier deep-self. According to the Argument from Tracing's historical condition, it too must be connected to an even earlier deep-self. And the same issue arises with this even earlier deep-self, and again with the ones before that, and so on, thus leading to an infinite regress.

⁴⁶ Ibid.; Mele, *Autonomous Agents: From Self-Control to Autonomy*.

⁴⁷ An account of moral responsibility that may potentially get around this problem, as helpfully pointed out by my examiner, is the following. *An agent A is MR for A at time t, only if A has the appropriate CAS at time t, unless the agent performed {actions} at times prior to t, such that A would have had the appropriate CAS at time, and the agent is MR for {actions} at times prior to t and {actions} at times prior to t were performed with the knowledge (at the respective times prior to t) that if they were performed there would be a high probability of the agent failing to have the appropriate CAS at time t.* However, I still think this account faces objections from section 2.4 and my arguments of Chapter 4.

Historical compatibilists now face a dilemma of how to end the regress. The first method for dissolving the regress involves discovering a point in time where the regress ends and we can say that this deep-self grants responsibility to all later deep-selves. The second strategy is to allow the infinite regress but deny that it is vicious or problematic for historical compatibilism.

The second horn of the dilemma cannot be solved by the historical compatibilist. Firstly, it is highly counterintuitive to think that our concept of moral responsibility requires tracing that stretches infinitely back in time. Such a position is not supported by any of our pre-theoretical intuitions, so it is highly unlikely that we would arrive at this position when we are in reflective equilibrium.⁴⁸ Moreover, even if our concept allows for linking deep-selves *ad infinitum*, historical compatibilists would have to conclude that no human being has ever been morally responsible. It is impossible for finite beings to complete an infinite series, so no human being would ever satisfy the regress. It is untenable for historical compatibilists to maintain that such a regress would not be vicious.

On the first horn of the dilemma, we must find a point in an agent's life where the regress terminates and ceases in such a way that it satisfies the historical compatibilist's necessary condition. The most obvious option would be to select the very first deep-self we had, which, by definition, cannot be linked to an earlier deep-self. This deep-self, which we most probably had at quite a young age, will still not fulfil the historical condition. This is because it leaves our moral responsibility ungrounded. The defender of the Argument from Tracing will have to view us as analogous to Beadie. Just as she was not responsible for her automobile accident because her actions cannot be traced to an earlier exercising of her

⁴⁸ The use of reflective equilibrium in philosophy was popularised in John Rawls, *A Theory of Justice* (Cambridge, Mass.: Harvard University Press, 1971).

agential powers, we would not be responsible for the products of our first deep-self because there is no earlier deep-self to anchor it.

Some philosophers believe that there is a solution to the first horn of this dilemma. Robert Kane, for example, believes that if the regress of actions that the agent is responsible for can terminate in the agent in such a way that the agent is the ultimate source of the actions, then the regress is not a threat to moral responsibility.⁴⁹ The solution applied to the regress at hand would involve discovering a time where the agent exercised her capacities for control, due to a combination of facts about her character and motives, where these facts explain entirely the exercising in question and the agent is responsible for these facts.

But this solution will not aid the historical compatibilist's defence of the Argument from Tracing for two reasons. First, there is scepticism about whether it is possible for any human person to satisfy Kane's sourcehood condition.⁵⁰ Second, even if our previous doubts are assuaged, this potential solution to the regress is necessarily incompatibilistic. It requires that the agent herself, and no facts outside the causal control of the agent, explain why the agent exercised her capacities for control when she did. But the Argument from Tracing is supposed to support a compatibilist account of moral responsibility. And compatibilists do not worry if the ultimate causal explanation of our behaviours can be traced all the way to the Big Bang, so long as we are appropriately invested in our actions when we perform them. Thus, this proposed answer to the regress contained in the Argument from Tracing will not assist the historical compatibilist, if they wish to remain a compatibilist.

The Argument from Tracing's necessary historical condition mirrors Strawson's argument for the impossibility of moral responsibility. That is, the conclusion of the Argument from Tracing holds that you are morally responsible for actions stemming from your compatibilist

⁴⁹ Robert Kane, *The Significance of Free Will* (New York: Oxford University Press, 1996).

⁵⁰ Pereboom, *Living without Free Will*, 41-50.

agential structure only if you are morally responsible for possessing that compatibilist agential structure, which requires a further compatibilist agential structure. This is similar to Strawson requiring you choose your character, and the character that chose that character, and so on.⁵¹ But historical compatibilists have explicitly rebuffed Strawson's ultimate control condition for being ridiculously over-demanding. Fischer, one of the main proponents of the Argument from Tracing, has described Strawson's view of moral responsibility as "a kind of metaphysical megalomania."⁵² As compatibilists reject Strawson's argument for requiring this kind of control, they should also reject the Argument from Tracing and its similarly over-demanding view of control.

For these reasons, I believe that the Argument from Tracing fails and should be found unconvincing by both sides of the debate.

2.4. Why the Argument from Tracing Does Not Support a Sufficient Historical Condition

My discussion of the Argument from Tracing could be objected to on the grounds that the argument was not intended to support the claim that history is necessary for moral responsibility, only that historical facts can be sufficient for moral responsibility. The tracing cases appear to demonstrate that Kima's past is sufficient to ground her responsibility for her crash, despite being drunk at the time it happened, rather than showing that these historical facts were necessary for her moral responsibility. It is important that non-historical compatibilists can reject this version of the argument, because if it can be shown

⁵¹ Galen Strawson, *Freedom and Belief* (Oxford: Clarendon Press, 1986).

⁵² John Martin Fischer, "The Cards That Are Dealt You," *The Journal of Ethics* 10, no. 1/2 (2006): 116.

that moral responsibility is historical in some areas, then it opens the door to the possibility that moral responsibility is historical elsewhere.

I do not believe that this version of the Argument from Tracing is successful. This is because moral responsibility in tracing cases is not genuinely historical. To demonstrate this, we need to examine why we believe agents can be held responsible in tracing examples. The reason why moral responsibility is preserved in the tracing cases is due to the fact that, before the agents incapacitate themselves, they could have foreseen or reasonably been expected to foresee what actions and omissions they would perform while incapacitated. To illustrate this point, imagine that Jimmy and Omar are walking in a good neighbourhood and Jimmy is persuaded by Omar to voluntarily jump into a hole that he cannot escape from on his own. While Jimmy is in the hole, Omar is assailed by a gang of muggers that requires at least two people to ward off.

Jimmy is in a comparable situation to individuals in tracing examples; he freely performed an action that led to him being incapacitated at a later time. Unlike the Kima example though, we do not think Jimmy is responsible for not being able to aid Omar. Jimmy is not blameworthy because he could not reasonably expect that Omar would be attacked while he was way down in the hole. Fischer & Ravizza, proponents of the Argument from Tracing, also accept that it is the fact of what one could expect or reasonably expect that explains our differing judgements of the agents in the tracing example.⁵³

When tracing examples are discussed, they are often presented so that it seems we start with an incapacitated agent performing an action and then trace the action back to a time where the agent was in control of what she was doing. Describing the examples in that manner makes it seem obvious that differences in history do matter to the agents' moral

⁵³ Fischer and Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, 195.

responsibility. But their responsibility is due to forward-looking factors; Kima is responsible because she should have expected that she would have a higher probability of crashing if she drunk and drove. As such, the tracing examples are being described back to front, misleading us into believing that it is the history of the agents that matter. A tracing example that accurately mapped on to our intuitions would begin with Kima, freely deciding to drink while being aware of the tragedy she could cause, and would then move forwards through time, checking to see if there is the right link between the agent now and the action they will perform at a later date.

The version of the Argument from Tracing currently being discussed hinges on there being a similarity between what the historical compatibilist wants to demonstrate and the tracing examples. Historical compatibilists believe we need to look back in time if we are to establish that a person is morally responsible for their actions. But, the tracing examples begin with an agent who we stipulate is morally responsible, and who then brings it about that they are not able to act freely at a later time. There is no parallel between the two because the tracing examples utilise a forward-looking phenomenon to be successful; they rely on the agent foreseeing or being expected to foresee the consequence of their current behaviour. Historical conditions on moral responsibility that historical compatibilists wish to defend are not similar to this. They do not refer to, or rely on an individual being able to foresee that they will become morally responsible for what they do. Instead, the historical compatibilists argue that an agent must see herself as an agent, or not have her capacities compelled in a certain way.⁵⁴

There is an established difference between tracing examples and the historical conditions that historical compatibilists defend. And the difference is such that there is no reason why

⁵⁴ Mele, *Autonomous Agents: From Self-Control to Autonomy*; Fischer and Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*.

acknowledging responsibility in tracing examples should lead a compatibilist to increase their credence in the truth of historical compatibilism. They are distinct phenomena, and so the Argument from Tracing is unable to support a historical sufficiency condition.

2.5. Conclusion

This chapter has contained reasons why the Argument from Tracing supports neither a necessary historical condition, nor a sufficient historical condition. This is because the argument that supports the necessary condition must conclude that no agent is ever morally responsible for what they do. Furthermore, tracing cases are unable to support either kind of historical condition because tracing is not genuinely historical, but a forward-looking phenomenon.

However, it is not problematic for my non-historical compatibilism if you do not buy the line of argument I have been advancing here. Instead, we can distinguish between cases of *direct moral responsibility* and *derivative moral responsibility*. You are directly morally responsible for an act if it issues from your compatibilist agential structure and other associated powers. You are derivatively morally responsible for an act if it can be appropriately traced back to an action that you were directly morally responsible for. The Argument from Tracing employs cases of derivative moral responsibility. If you still hold that the Argument from Tracing shows that moral responsibility is historical in some sense, then we can say it is historical in that sense for actions for which you are derivatively morally responsible. The rest of this thesis however, will be concerned with direct moral responsibility and whether that is genuinely historical. Specifically, in Chapter III I will give a detailed explanation of the historical compatibilist's manipulation argument, and then

provide arguments against it in Chapter IV. For any defenders of the Argument from Tracing, assume that succeeding references to moral responsibility are cases of direct moral responsibility.⁵⁵

⁵⁵ This focus on direct moral responsibility can be found in McKenna, "Moral Responsibility, Manipulation Arguments, and History: Assessing the Resilience of Nonhistorical Compatibilism," 155-57.

Chapter III – Explaining Manipulation Arguments for Historical Compatibilism

3.1. Introduction

Chapter II focused on the Argument from Tracing, which was supposed to show that moral responsibility is historical. Though this argument fails, historical compatibilists are able to rely on a second argument; the manipulation argument. This chapter will expand upon what a manipulation argument involves, though we have already encountered a manipulation argument when discussing Locke's objection to Frankfurt's non-historical compatibilist position. Manipulation arguments are a popular dialectical move, originally used by incompatibilists to demonstrate the falsity of compatibilism, though they can be employed in defence of historical compatibilism.⁵⁶ I will be careful to distinguish between manipulation arguments for incompatibilism and manipulation arguments for historical compatibilism.

In Section 3.2 of this chapter, I will specify what is meant when philosophers talk about manipulation in this debate, and Section 3.3 will highlight the troublesome features of using the word 'manipulation.' Section 3.4 will make distinctions between kinds of manipulation arguments. Section 3.5 will present an example of a manipulation argument pertinent to our

⁵⁶ Incompatibilist manipulation arguments can be found in Robert Kane, *Free Will and Values* (Albany: State University of New York Press, 1985), 34-49; John Martin Fischer, "Excerpts from John Martin Fischer's Discussion with Members of the Audience," *Journal of Ethics* 4, no. 4 (2000): 414-18; Joshua Greene and Jonathan Cohen, "For the Law, Neuroscience Changes Nothing and Everything," *Philosophical Transactions of the Royal Society of London: Biological Sciences* 359, no. 1451 (2004): 1780; Locke, "Three Concepts of Free Action," 104-06; Richard Taylor, *Metaphysics*, Fourth ed. (Englewood Cliffs, N.J.: Prentice-Hall, 1974), 43-44; Ishtiyaque Haji, *Moral Appraisability* (New York: Oxford University Press, 1998), 108. Derk Pereboom's incompatibilist Four-Case Argument, which I shall discuss, can be found in Derk Pereboom, "Determinism Al Dente," *Noûs* 29, no. 1 (1995); *Living without Free Will*, 100-26; *Free Will, Agency, and Meaning in Life* (Oxford: Oxford University Press, 2014), 74-79. Incompatibilist manipulation is also present in Alfred R. Mele's Zygote Argument, which can be found in Mele, *Free Will and Luck*, 184-95; "Manipulation, Compatibilism, and Moral Responsibility," 278-84; "Moral Responsibility and Agents' Histories," 179, footnote 24; "Manipulation, Moral Responsibility, and Bullet Biting," *The Journal of Ethics* 17, no. 3 (2013): 175-83.

interests and Section 3.6 will examine the structure of this argument. Finally, in Section 3.7, I will investigate reasons why this argument has been considered powerful.

3.2. The Nature of Manipulation

The manipulation philosophers are concerned with always contains some essential elements, regardless of whether the manipulation is being used to defend historical compatibilism or incompatibilism. However, the manipulation is very different to the garden-variety manipulation that we commonly refer to in ordinary life. For example, we normally think of manipulated agents as beings who have been deceived about what is actually the case and are misled into carrying out an act they would rather not perform. The manipulation we will discuss will not cause the agents to have false beliefs that are relevant to the ends they are attempting to realise (though they may be deceived about the true origin of their ends, but this is a fact that could be true of all of us). Our manipulated agents have correct beliefs about what they desire and the consequences of their actions. Moreover, unlike Susan Wolf's example of a manipulated agent Jojo, the manipulated agents we will focus on have not been rendered insane and understand moral reasons perfectly well.⁵⁷

When we imagine manipulating someone to do our bidding, we often picture the use of force and coercion. The manipulated individual is held at gun point and constrained to do something they would 'deep down' not want to do. This is not the case with the philosophically interesting kind of manipulation. The manipulation we will discuss involves the removing of previous wants and values that the actor had, and replacing them with a

⁵⁷ Wolf, "Sanity and the Metaphysics of Responsibility."

new set of wants and values. Via this method, the manipulator ensures that he will get his desired outcome by creating a new system of motives that cause the manipulated actor to act as he wants. The victim's new motives both conform to what the manipulator wants, and what she now regards as the best course of action for achieving her new purposes. As Kane notes, when this manipulation occurs, "the controlled agent's will is not frustrated. No conditions prevent the agent's wants, desires, and intentions from being realised in action."⁵⁸

It is important to stress these points because the purpose of the manipulation arguments is to draw an analogue between the manipulated agent and a normal agent exercising control who lives in a deterministic world. The philosophically interesting manipulation that we will discuss is such that when it occurs, it is the case that the manipulated agent satisfies a plurality of the reputedly best compatibilist control conditions. For the manipulation arguments to succeed, they must as closely approximate the non-historical compatibilist's criteria for control as they can. If the manipulation is under-described so that, for example, the manipulation failed to give an agent a coherent deep-self, then we would be able to automatically reject manipulation arguments for failing to point out a similarity with normal humans.

A striking example of the philosophically interesting kind of manipulation is given in B.F. Skinner's *Walden Two*. Walden Two is a utopia where the community has been conditioned to possess a certain set of desires. The founder of Walden Two, Frazier, describes the freedom of the inhabitants of his purported paradise in the following:

⁵⁸ Kane, *Free Will and Values*, 34.

We can achieve a sort of control under which the controlled, though they are following a code much more scrupulously than was ever the case under the old system, nevertheless *feel free*. They are doing what they want to do, not what they are forced to do. That's the source of the tremendous power of positive reinforcement – there's no restraint and no revolt. By a careful cultural design, we control not the final behaviour, but the *inclination* to behave – the motives, the desires, the wishes.⁵⁹

All that happens is contained in an original plan, yet at every stage the individual seems to be making choices and determining the outcome. The same is true of Walden Two. Our members are practically always doing what they want to do – what they 'choose' to do – but we see to it that they will want to do precisely the things which are best for themselves and the community. Their behaviour is determined, yet they're free.⁶⁰

The citizens are much like us in that they are doing what they want to do, but what they want to do is due to the conditioning of their society. This is the kind of manipulation that will be dealt with in this essay.

3.3. Loaded Terminology

Before I turn to manipulation arguments, I wish to note that the word 'manipulation' has many negative connotations that are extraneous to whether we should view the arguments as successful. In the literature, a worry has been raised by philosophers discussed below, that the efficacy of the intuition behind manipulation arguments is in part due to the meaning of how the word is normally used. That is, the negative connotations of 'manipulation' subtly cause us to feel that the manipulation argument is more powerful than we ought to believe.

Dana Kay Nelkin notes that real world manipulation often involves automatic behaviour that is insensitive to reasons.⁶¹ The manipulated agents we will consider will be reasons-responsive, capable of recognising and responding to reasons in actual and counterfactual

⁵⁹ B.F. Skinner, *Walden Two* (New York: Macmillan, 1958), 218.

⁶⁰ *Ibid.*, 247.

⁶¹ Dana Kay Nelkin, *Making Sense of Freedom and Responsibility* (Oxford: Oxford University Press, 2011), 56, 58.

scenarios. David Blumenfield points out that “mind controller” has a “ghoulish ring” to it.⁶² This makes it sound like the “victim” is to be dominated in such a way that they act only on impulses that are discordant with who they truly are, which have been ingrained in them by the mind controller, similar to the manipulation in *A Clockwork Orange* or *1984*.⁶³ But as stated earlier, the manipulated agent will satisfy various compatibilist theories, including the deep-self view.

Bruce Waller states that the fear of being controlled by a neuroscientist who is exploiting us “*is doing a lot of the work*” in the manipulation argument. And by comparing the manipulated agent and normal agent, we come to see the normal agent as a victim of an environment with ulterior motives. Waller argues that this view is mistaken because the environment has shaped our species’ evolution so that we can thrive in it.⁶⁴ Waller, like Nelkin and Blumenfield, believes that the vocabulary we adopt influences how we perceive the manipulation argument.⁶⁵

It seems that there is a real risk that ‘manipulation’ is a loaded term that subconsciously biases us against non-historical compatibilism. However, I will continue to use the term ‘manipulation’ for two reasons. The first is that it is commonly used in the literature. Secondly, as part of the principle of charity, granting my opponents use of the word ‘manipulation’ allows them to construct the strongest possible arguments. It also strengthens my position as I believe that non-historical compatibilism can be seen as intuitive despite the implications of the words used. However, I do ask that the reader keep

⁶² David Blumenfeld, "Freedom and Mind Control," *American Philosophical Quarterly* 25, no. 3 (1988): 224.

⁶³ Anthony Burgess, *A Clockwork Orange* (New York: Heinemann, 1962); George Orwell, *1984* (New York: Secker & Warburg, 1949).

⁶⁴ Bruce N. Waller, "Free Will Gone out of Control," *Behaviourism* 16, no. 2 (1988): 155-56.

⁶⁵ Daniel Dennett makes a similar point in Daniel Dennett, *Elbow Room: The Varieties of Free Will Worth Wanting* (Cambridge, MA: The M.I.T. Press, 1984), 10.

in mind the specific meaning 'manipulation' has in this context when reading the arguments.

3.4. Global and Local Manipulation⁶⁶

The manipulation argument for incompatibilism and manipulation argument for historical compatibilism utilise different kinds of manipulation to establish their conclusion. The first kind of manipulation is what I shall call *global manipulation* and it features in manipulation arguments for incompatibilism. Global manipulation refers to manipulation that occurs before the manipulated agent is able to develop their own character, such as when they are neonatal or in utero. Global manipulation affects the trajectory of the agent's life so that they come to adopt a specific set of preferences, whilst having as much control over their decisions as anyone who possesses a compatibilist agential structure. Walden Two is a case of global manipulation because its citizens are conditioned from an early age, before they can independently develop their own preferences.

Global manipulation is employed by incompatibilists to target compatibilism. The idea is that a globally manipulated agent is no different to a person born into a world where metaphysical determinism is true; the difference is only that global manipulation is artificial. Incompatibilists believe that it is intuitive that a globally manipulated agent is not responsible, because it is the brainwashers, not the agent, who are the true source of the behaviour. Since there is no difference between a globally manipulated agent and a determined agent, incompatibilists conclude that determined agents are not responsible either, so compatibilism is false. Consider Walden Two. It is possible that a normal agent in a

⁶⁶ My use of the terms global and local manipulation differ from its usage elsewhere in the literature, such as in McKenna, "Responsibility and Globally Manipulated Agents."

deterministic universe could satisfy identical compatibilist conditions of control to those of a citizen of Walden Two. And if we accept the incompatibilist's premise that the citizens of Walden Two are not responsible because they have been conditioned by forces outside of their control, then on pain of contradiction, we should accept that normal agents in deterministic universes are not responsible either and that compatibilism is false.

All compatibilists who are party to the historical debate agree that this global manipulation argument fails. We must think the same if we are assuming compatibilism is true and wish to engage in the historical debate. The manipulation we are interested in is *local manipulation*, which refers to manipulation that has occurred after the agent has cultivated their own pro-attitudes. In a typical local manipulation case, an agent who has their own values and desires will have these existing pro-attitudes obliterated and exchanged with a completely new set of pro-attitudes. Agents who are locally manipulated stand in contrast to globally manipulated agents, as locally manipulated agents are normally adults who have already developed a deep-self, then lose it, and then are implanted with another, whereas global manipulation operates so that the agent's original deep-self is the deep-self that is created by brainwashing. As we shall see, Mele's Chuck and Beth case is an example of local manipulation.⁶⁷

We can partially define the important philosophical positions relevant to this subject in reference to how they view the different kinds of manipulation as affecting free will and moral responsibility. Non-historical compatibilism and incompatibilism both hold that you

⁶⁷ Arguments involving local manipulation can be found in Michael A. Slote, "Understanding Free Will," *The Journal of Philosophy* 77, no. 3 (1980): 149-50; Eleonore Stump, "Control and Causal Determinism," in *Contours of Agency: Essays on Themes from Harry Frankfurt*, ed. Sarah Buss and Lee Overton (Cambridge, Mass.: MIT Press, 2002), 47-49; Fischer and Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, 230-36 and Mele, *Autonomous Agents: From Self-Control to Autonomy*, 144-73; "Manipulation, Compatibilism, and Moral Responsibility," 266-75; "Moral Responsibility and History Revisited," 464-65; "Moral Responsibility and Agents' Histories," 166-68; "Moral Responsibility, Manipulation, and Minutelings," *The Journal of Ethics* 17, no. 3 (2013): 153-54; "Manipulation, Moral Responsibility, and Bullet Biting," 169-70.

should give symmetrical judgements to global and local manipulation cases. Incompatibilists believe that both global and local manipulation rob you of your freedom as the agent is no longer the appropriate source of their behaviour. Non-historical compatibilists deny this and state that the sourcehood requirement can be satisfied regardless of whether those kinds of manipulation have occurred. In Chapter IV, I will defend the claim that historical compatibilism is committed to an asymmetry in judgement between the two kinds of manipulation. For now, I will simply state that historical compatibilists believe local manipulation nullifies freedom and moral responsibility, but do not view global manipulation as a threat to either free will or moral responsibility.⁶⁸

3.5. Mele's Local Manipulation Case

Now I will present Mele's case of local manipulation. I believe it is a particularly clear example of the kind of manipulation used in manipulation arguments by historical compatibilists against non-historical compatibilists. I will explicate the details of the case and the intuitions involved, and in Section 3.6 I will discuss how it fits into the manipulation argument for historical compatibilism.

Imagine a despicable agent Chuck. Chuck has always believed that morality is a system of rules used by the weak to bind the strong. He has always valued being the kind of person who can throw off the shackles of morality and pursue his own egoistic ends. Unfortunately for Chuck, he is still empathetic and experiences pangs of guilt when he performs wrong acts, like torturing animals. To overcome his perceived weaknesses, Chuck intentionally sets

⁶⁸ Theoretically, there could be a position that holds the opposite of historical compatibilism, and views global manipulation as problematic but not local manipulation. As no one has currently defended the position in the literature, I will not discuss it here.

out to perform harsher and crueller acts more frequently, such as bullying weaker and more vulnerable people than himself, so he may rid himself of his guilt. Over time, he no longer feels any remorse for what he does and begins to take pleasure in grisly and violent acts, which culminates in Chuck murdering his neighbour.

Now consider Beth. Beth is a saint who constantly strives to do what is morally right and thoughts of violence would never cross her mind. However, during the night a team of neuroscientists, who have discovered the system of values that causes Chuck to act violently, implant the same system of values in Beth while erasing her previous values. They also leave her memories intact.⁶⁹ When she wakes in the morning, she is surprised to find that she would really like to murder her neighbour. As she reflects on this desire, she judges, in line with her new values, that morality is just a constraint to her achieving her new ends. As it conforms to her new values and she would really like to do it, Beth hunts down and kills her neighbour Frank.⁷⁰

Both Beth and Chuck are doing what they want and intend to do when they murder their respective neighbours. But intuitively, only Chuck is blameworthy for what he does. The values Beth acts from do not seem to belong to her in the sense required by moral responsibility. This is the intuition that historical compatibilists are trying to engender from their local manipulation cases. I will now explore how this thought experiment supports the historical compatibilist's local manipulation argument.

⁶⁹ This aspect of the thought experiment is in anticipation of the objection that Beth is no longer the same person after the manipulation.

⁷⁰ This example comes from Mele, *Free Will and Luck*, 171-74. Similar cases first appeared in *Autonomous Agents: From Self-Control to Autonomy*, 144-73.

3.6. The Structure of a Manipulation Argument for Historical Compatibilism

To critically evaluate a manipulation argument, we first need to understand its structure. A manipulation argument combines three insights. The first comes from the example of local manipulation we encountered in Section 3.5; when Beth is subjected to clandestine manipulation, it is intuitive that she is not morally responsible for actions that are the direct product of this manipulation, actions that she would not have performed had it not been for the manipulation. The second intuition that we have is that non-manipulated agents, like Chuck, can be legitimately blamed or praised for what they do, though they may perform the same acts for the same reasons as the manipulated agents. Lastly, we reasonably believe that there are no snapshot properties that can explain the difference in moral responsibility between agents like Beth and Chuck.

The general structure of a manipulation argument for historical compatibilism is as follows:

1. For any agent *S*, if *S* is, like Beth, locally manipulated in manner *X* to perform action *A* from her compatibilist agential structure *CAS*, then *S* does not *A* of her own free will and is therefore not morally responsible.
2. For any normally functioning agent *F*, like Chuck, who *As* from *CAS*, does *A* of his own free will and is therefore morally responsible.
3. There is no difference between *S* and *F*, in terms of their ahistorical *CAS*'s, which can explain the difference in their moral responsibility for *A*.

4. Therefore, satisfying an ahistorical CAS is insufficient for free will and moral responsibility. There is a further historical requirement.⁷¹

The conclusion is a statement of historical compatibilism, though this argument is agnostic as to whether it will be a positive or negative historical condition. Mele has used the argument to defend the position that you are not morally responsible if you involuntarily acquire your values via a process that bypasses your rational capacities and these new values are not supported by any already existing values the agent may have. In contrast, Fischer & Ravizza would assert that the argument shows that agents need to have a history whereby they take responsibility for the mechanism on which they act in order to be morally responsible agents.

Though they disagree about why it is the case, all historical compatibilists unite on the point that local manipulation places an agent outside the scope of moral responsibility. They thus agree that the above argument is sound. To resist historicism, non-historical compatibilists must find a fault with one of the premises. It does not seem helpful to deny premises (2) or (3). Premise (2) simply suggests that ordinary agents who satisfy compatibilist conditions for free will are in fact free. To deny (2) would make one an incompatibilist. Premise (3) states that both a locally manipulated agent and an ordinary agent can have the same properties that constitute the best compatibilist conditions of moral responsibility. I do not see any plausibility in denying (3); it seems that any compatibilist account of moral responsibility that can be brought about in a naturally

⁷¹ For formal statements of global manipulation arguments, see Tomis Kapitan, "Autonomy and Manipulated Freedom," *Philosophical Perspectives* 34, no. 14 (2000): 85-86; Kadri Vihvelin, "Arguments for Incompatibilism," ed. Edward N. Zalta, *The Stanford Encyclopedia of Philosophy* (<http://plato.stanford.edu/archives/spr2011/entries/incompatibilism-arguments/2007>); Michael McKenna, "A Hard-Line Reply to Pereboom's Four-Case Manipulation Argument," *Philosophy and Phenomenological Research* 77, no. 1 (2008): 143.

deterministic world can be brought about artificially. There is no reason to claim that a manipulated agent could not have a deep-self but a normally functioning agent can have a deep-self.

This leaves only premise (1) to deny. A non-historical compatibilist must find convincing reasons to make it seem plausible that manipulated Beth is in fact, contrary to our initial intuition, free and responsible for the murder she committed. This is no easy feat given the strength of the historicists' intuition. In Chapter IV I will take up this task.

3.7. The Strength of Manipulation Arguments

The reaction that I expect to be engendered by Mele's manipulation argument above is one that denies Beth is morally responsible, though Chuck is responsible. Instinctually, there is something about manipulation that confers the status of non-responsibility to those who undergo it. In this section, I will examine theories that endeavour to explain what it is about local manipulation that causes historical compatibilists to think that it nullifies moral responsibility.

To be a successful explanation, it must achieve three things. Firstly, it must point out the property that a local manipulation case has and explain why we believe it counteracts moral responsibility. Secondly, the property identified must be shared by all instances of manipulation found problematic by the theorist. For historical compatibilists, this property must be found in all cases of *local* manipulation. Thirdly, this property cannot be shared by other phenomena we do not think conflict with moral responsibility; it must be unique to only the intuitive cases. To satisfy this aim, historical compatibilists' cannot single out a

property that is also found in, for example, global manipulation cases, because historical compatibilists do not think they rule out moral responsibility.

The first two theories I will consider and then rebuff for failing to fulfil the three desiderata involve the Transfer Principle, discussed in Section 3.7.1, and Mele's radical reversal suggestion, debated in Section 3.7.2. Finally, in Section 3.7.3, I will endorse the last remaining explanation that I have found in the literature, presented by Fischer & Ravizza. They claim that it is simply intuitive that manipulation and similar kinds of cases undermine moral responsibility.

3.7.1. Their Strength as Derived from the Transfer Principle

One method of supporting the manipulation argument rests on the controversial philosophical principle, the Transfer Principle. This principle states that if p obtains and no one is responsible for p , and if p brings about q and no one is responsible for this fact either, then no one is morally responsible for q obtaining.⁷² More simply, the Transfer Principle holds that an agent cannot be morally responsible for an action if it is the deductive entailment from a state of affairs for which the agent is not morally responsible. In applying the principle to the present argument, the historical compatibilist might argue that if an agent is not responsible for the fact that she was manipulated, and this manipulation caused her to perform an action, then the agent is not responsible for that action.

Although this might seem promising, many compatibilists reject the Transfer Principle, including Fischer & Ravizza. One problem is that the principle has apparent

⁷² For a statement of the transfer principle, see Inwagen, *An Essay on Free Will*, 182-88.

counterexamples, such as in cases of overdetermination.⁷³ Suppose that Betty intends to cause an avalanche that will crush a camp at the bottom of a mountain. She sets off her dynamite at t_1 , and the camp is destroyed by the avalanche at t_3 . Unbeknownst to Betty, the mountain she is on has been slowly eroding over the years. The erosion is so severe that if Betty had not caused an explosion at t_1 , a natural avalanche would have formed at t_2 , which would have squashed the camp at t_3 . According to the Transfer Principle, since no one is responsible for the naturally occurring erosion, and no one is responsible for the fact that if the erosion continues, there will be an avalanche that obliterates the camp at t_3 , then it is the case that no one is responsible for the fact that an avalanche obliterated the camp at t_3 . But this is false. Betty is responsible for the avalanche that demolished the camp. Thus, defenders of the Transfer Principle must overcome these problems if it is to be adopted by historical compatibilists.⁷⁴

A further reason why compatibilists would be motivated to reject the Transfer Principle is that it is a decidedly incompatibilistic principle. It is a central component of two arguments for incompatibilism; the Consequence Argument and the Direct Argument.⁷⁵ I do not have space to expand on these arguments, but I will note that one method compatibilists adopt to combat these arguments is to deny the Transfer Principle, so some compatibilists are unlikely to endorse the principle. Furthermore, it enshrines an incompatibilistic tendency. Incompatibilists often believe that if we are to be responsible for an act, we must also be

⁷³ For the original version of the example, as well as an excellent discussion of the Transfer Principle, see Fischer and Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, 151-69; Mark Ravizza, "Semi-Compatibilism and the Transfer of Non-Responsibility," *Philosophical Studies* 75, no. 1/2 (1994).

⁷⁴ Contemporary philosophers continue to disagree over the status of transfer principles. For example, a defence of a new transfer principle can be found in Ted A. Warfield, "Determinism and Moral Responsibility Are Incompatible," *Philosophical Topics* 24, no. 2 (1996). For a criticism of that principle, see Eleonore Stump and John Martin Fischer, "Transfer Principles and Moral Responsibility," *Philosophical Perspectives* 34, no. 14 (2000).

⁷⁵ For information on these two arguments, see Fischer and Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, 17-24, 151-54.

responsible for any conditions sufficient for that act's occurred. Compatibilists accept that there can be a long deterministic chain of events that precede our action, yet this does not affect whether we are responsible or not for our actions. As they are historical *compatibilists*, their appeals to manipulation arguments do not rest on the Transfer Principle.⁷⁶

3.7.2. Mele's "Radical Reversal" Suggestion

Mele has proposed an explanation of why manipulation eliminates moral responsibility, what he has dubbed the *radical reversal* suggestion. Elaborating, the suggestion is:

Most lay readers of Beth's story would suppose that (at least prior to her radical transformation) she was a morally responsible agent and that she was morally responsible, at least to a significant extent, for having the character – or collection of values – she had. I suppose this too. Indeed, I make this a feature of the story. Given this feature, a more interesting and attractive suggestion is ready to hand: Beth's pre-transformation character was sufficiently good that killing [Frank] was *not even an option for her*; and the combination of this fact with the fact that Beth was morally responsible (to some significant degree) for that character, facts about her history that account for her moral responsibility for that character, and the facts that account for her killing [Frank] suffices for her not being morally responsible for killing him. When I ask myself why my gut reaction to Beth's story (insofar as I can have a gut reaction to it) is what it is, these are the considerations that loom large. (Emphasis Mele's)⁷⁷

I have reproduced Mele's quote in full because I am uncertain of my interpretation of this passage and do not wish to be seen as offering an uncharitable reading. I believe that Mele is pointing to two different sets of descriptive facts to explain why locally manipulated agents are not morally responsible. One set is the collection of facts regarding Beth's history, which include the facts about her old character, the fact she was manipulated, and facts about her new character and how it figures into the murder of Frank. The second

⁷⁶ A similar point is made in Gary Watson, "Reason and Responsibility," *Ethics* 111, no. 2 (2001): 386.

⁷⁷ Mele, "Moral Responsibility and Agents' Histories," 169.

collection of facts Mele cites refers to the contrast between Beth's pre-manipulation nature and post-manipulation nature. The second set of facts account for the *radical reversal* aspect of the manipulation. Judging from Mele's use of emphasis, a manipulative reversal of someone's character is radical when, with their new character, the manipulated agent is willing to perform actions that would not have been an option for them with their pre-manipulated character. According to Mele, what explains the intuitions in manipulation cases is that the newly engineered characters are radically different to the recently removed characters.

Mele's explanation is appealing, and is consonant with intuitions with other cases he presents. Imagine that Dennis and Dee are ordinary human beings who have been allocated a sizeable fortune from their mother's will. Dee was disliked by her mother, so the will is set up such that Dee will only receive her portion once Dennis has passed, provided that Dennis has not spent the entirety of the allotment. Dee believes it is wrong for her to murder Dennis, but does so for financial gain. Now imagine that Beth, who used to be kind-hearted but has been manipulated to endorse all of Chuck's psychopathic values, is also in the same position as Dee with respect to her brother Frank. Beth believes it is wrong to kill for financial gains, but due to her manipulation, she is glad for the opportunity to murder Frank. Dee and Beth both satisfy a large set of strict compatibilist control conditions.

Additionally, imagine a being Mele called a *minuteling*. A minuteling is a cloned human being that has been unconscious for years, and has a bunch of pseudo-memories that match the original's memories. Limits of the technology mean that a minuteling can only survive for a minute.⁷⁸ A minuteling Mac is brought to life to examine what the real Mac would do, as Mac's brother Charlie has also received financial priority in their mother's will. Minuteling

⁷⁸ "Moral Responsibility, Manipulation, and Minutelings," 162.

Mac finds himself in the same circumstances as Dee and Beth, with respect to his minuteling brother Charlie (who he falsely remembers and believes to be his actual non-minuteling brother). Though he knows it is wrong, Mac murders Charlie. Furthermore, minuteling Mac exercises the exact degree of compatibilist control that Dee and Beth have.

Mele believes that Dee is responsible for the murder she commits, that Beth is not, and that neither of these beliefs commits him to a claim about the responsibility or non-responsibility of Mac.⁷⁹ Dee is responsible because she has the right kind of history and her action conforms to the kind of person she is. Beth however, has been radically reversed so Mele thinks she is not responsible. Minuteling Mac's act of murder is not at odds with who he 'really' is, so Mele is able to either claim that he is responsible for his act, or he can fish around for another reason to say Mac is not responsible.⁸⁰ That the radical reversal suggestion can account for these varied cases is a credit to the view.

Mele's argument is that facts about Beth's history as well as her reversal of character are sufficient for Beth deserving no blame for the murder she undertakes. However, I cannot understand why someone with historical compatibilist sympathies would agree that these two conditions are *jointly* sufficient. I believe that it can be seen that the radical reversal of character condition is redundant, and that historical compatibilists will have to rely on Beth's history of manipulation to carry all of the intuitive weight of their argument.

Suppose that Liz prefers to eat smooth peanut butter to crunchy peanut butter to the point that she rarely opts for a crunchy peanut butter sandwich if the other spread is available. In an effort to boost sales, the crunchy peanut butter lobby hire a neuroscientist to manipulate Liz's preferences so that she now desires crunchy peanut butter over smooth

⁷⁹ Ibid., 163-64.

⁸⁰ I believe our intuitive reaction to beings like minutelings is that they are responsible for what they do. I will return to the topic of ahistorical agents' moral responsibility later.

peanut butter. When she next goes to make a sandwich, she reflects on her new preferences and, despite the availability of smooth peanut butter, she chooses to have a crunchy peanut butter sandwich. Both prior to and after the manipulation, Liz fulfils the correct compatibilist control condition.

If you do not believe Beth freely chose and is responsible for the murder she commits, then for similar reasons, you should not believe that Liz freely chose and is responsible for her lunch decision. It is highly unlikely that Liz would have chosen crunchy peanut butter, had it not been for the manipulation. If one accepts this line of reasoning, then Liz's case becomes a counterexample to Mele's radical reversal suggestion. Liz is not radically reversed, on my analysis of Mele, because prior to the manipulation, there were situations, albeit rare ones, in which Liz would eat a crunchy peanut sandwich. Crunchy peanut butter was an option for Liz both before and after the manipulation, yet we still find that she is not morally responsible for decisions involving her newly implanted pro-attitude.

It is possible that Mele may object to this counterexample on the grounds that murder is an action which is very clearly morally wrong, whereas eating crunchy peanut butter is a morally neutral action and that the difference in the action's moral status is affecting our intuitions. It could be that our evaluations of the different kind of acts are influencing our judgements of the agent's responsibility. However, I do not think this reply will succeed. Even if Mele can convincingly argue that Liz is manipulated, yet still responsible, because she performed a neutral act, I do not think this will protect his radical reversal theory from counterexamples. This is because we could put forward a new thought experiment, where a villainous character is manipulated so that they will be more certain to perform very morally wrong actions, though performing these actions was not completely out of the question for their old character. This would be a thought experiment where an agent's character is

revised, but not radically, and the action they perform is not neutral, thus getting around Mele's objection. I believe that this thought experiment would show that non-radically revised agents who perform morally wrong actions are also not responsible for what they do.

Both Liz and Beth were intentionally manipulated, so that they would have a set of values that would result in them performing a particular action. It is this fact, and not a fact about the contrast between their new character and their old character, that pumps the intuition that they are not morally responsible for what they do. Though cases of radical reversal may make historical compatibilist intuitions more salient, I do not see why responsibility-undermining manipulation is restricted only to those cases. Liz's conservative reversal of character should be regarded as equally responsibility nullifying. For this reason, I reject Mele's radical reversal suggestion as an explanation of what it is about manipulation cases that disposes us to resist non-historical compatibilism.

3.7.3. *Intuitions and Manipulation Arguments*

Instead of relying on a principle, or Mele's radical reversal suggestion, the force of manipulation arguments is to come from the fact that we view cases of manipulation as involving "intuitively responsibility undermining factors."⁸¹ We are simply meant to intuitively perceive that manipulation stands in contrast with being morally responsible. Accepting this strategy means abandoning the project to offer reasons for what makes manipulation problematic, at least for the present, and instead focuses our attention on finding the position that covers the greatest number of intuitions possible. Given the failure

⁸¹ Fischer and Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, 202.

of the past two attempts at systematising our thoughts on manipulation, I believe historical compatibilists should embrace this procedure.

Furthermore, it appears historical compatibilists' current practice reflects the fact that they are simply trying to give an account that best captures our intuitions, as opposed to one which can be defended on the basis of a principle. For example, Mele argues that the significant difference between Chuck and Beth is that Chuck's "values were acquired under [his] own steam, whereas Beth's were imposed upon her."⁸² Mele seems to be saying that having a history of acquiring values under your own steam is what explains the difference between responsible and non-responsible agents. But such a distinction is a *positive* historical constraint; it states that an agent *must* have a history of acquiring values under their own steam to be considered responsible. Although Mele appears to accept that this explains what is wrong with manipulation, he then goes on to endorse a *negative* historical constraint, which allows him to hold that beings like minutelings could be morally responsible. This suggests that rather than trying to give an account of what is problematic with manipulation, he is content with providing the most intuitive theory.⁸³

I believe that this explicit reliance on intuitions does not make the historical compatibilist's argument any less cogent. However, I am aware that there is a debate among philosophers over whether it is appropriate to use intuitions at all in philosophical arguments. That debate is beyond the scope of this thesis. I will simply note that a number of philosophers defend the role intuitions play and assume they are correct, like many other

⁸² Mele, *Autonomous Agents: From Self-Control to Autonomy*, 155.

⁸³ For arguments that historical compatibilists can only defend positive historical requirements, see Benjamin Matheson, "Compatibilism and Personal Identity," *Philosophical Studies* 170, no. 2 (2014): 325-26; McKenna, "Moral Responsibility, Manipulation Arguments, and History: Assessing the Resilience of Nonhistorical Compatibilism," 167-69. For objections to these kinds of arguments, see Ishtiyaque Haji, "Historicism, Non-Historicism, or a Mix?," *The Journal of Ethics* 17, no. 3 (2013): 196-204.

philosophers implicitly do.⁸⁴ But, knowing that it is intuition that the argument derives its persuasive power from supplies non-historical compatibilists with a strategy for dealing with manipulation arguments. Firstly, non-historical compatibilists must undermine the intuition that manipulation is responsibility undermining. This will ameliorate the counterintuitive aspect of non-historical compatibilism. Secondly, non-historical compatibilists must show that it is in fact their position which is best supported by our intuitions. This will involve presenting cases where it is the historical compatibilists who give an absurd response and where non-historical compatibilism gets it right.⁸⁵

3.8. Conclusion

In this chapter, I have explored and defined conceptual issues related to the manipulation argument and laid the groundwork so that we have a framework for how best to criticise the argument. Specifically, in Section 3.7 I argued that what motivates the manipulation argument is simply the intuition that manipulated agents are neither free nor responsible for what their actions. In Chapter IV, I will debate this point by demonstrating that it is not obvious that manipulated agents are unfree, that historical compatibilism has

⁸⁴ For philosophical arguments defending the role of intuitions in argument, see George Bealer, "The Incoherence of Empiricism," *Proceedings of the Aristotelian Society* 66 (1992); Joel Pust, "Against Explanationist Scepticism Regarding Philosophical Intuitions," *Philosophical Studies* 106, no. 3 (2001); Alvin I. Goldman, "Philosophical Intuitions: Their Target, Their Source, and Their Epistemic Status," *Philosophische Studien* 74 (2007); Jennifer Nagel, "Intuitions and Experiments: A Defense of the Case Method in Epistemology," *Philosophy and Phenomenological Research* 85, no. 3 (2012). For arguments against the use of intuitions, see Robert Cummins, "Reflections on Reflective Equilibrium," in *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*, ed. Michael Raymond DePaul and William M. Ramsey (Lanham: Rowman & Littlefield, 1998); Stacey Swain, Joshua Alexander, and Jonathan M. Weinberg, "The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp," *Philosophy and Phenomenological Research* 76, no. 1 (2008); Jonathan M. Weinberg, Haun Nichols, and Stephen Stich, "Normativity and Epistemic Intuitions," *Philosophical Topics* 29, no. 1/2 (2001); Joshua Knobe and Shaun Nichols, "An Experimental Philosophy Manifesto," in *Experimental Philosophy*, ed. Joshua Knobe and Shaun Nichols (New York: Oxford University Press, 2008).

⁸⁵ Emphasising the role intuitions play in the manipulation argument comes from Watson, "Reason and Responsibility," 386.

counterintuitive implications, and that non-historical compatibilism captures many intuitions that historical compatibilism is unable to.

Chapter IV – Criticising Manipulation Arguments for Historical Compatibilism

4.1. Introduction

From Chapter III, it is clear that for non-historical compatibilism to be a palatable position, at least one of two claims must be defended. Either the non-historical compatibilist needs to show that it is not contrary to our intuitions to think locally manipulated agents are morally responsible for what they do, or that we have overall more reasons to believe this statement, despite finding it counterintuitive. In this chapter, I will be arguing for both these claims by focusing primarily on the local manipulation argument. I believe this is crucial for a defence of non-historical compatibilism, because I believe it is the local manipulation argument that motivates most compatibilists to affirm a historical constraint.

In Section 4.2, I will address replies to manipulation arguments that may tempt non-historical compatibilists but I believe are ineffectual. The remaining sections will examine arguments that are more promising. Sections 4.3 and 4.5 will undermine the reasons we have for accepting historical compatibilism, and will provide a normative reason for accepting non-historical compatibilism. Sections 4.4, 4.6, 4.7, and 4.8 will show that it is in fact intuitive that locally manipulated agents are morally responsible for what they do. Sections 4.9, 4.10, and 4.11 will argue that locally manipulated agents are not relevantly different to certain kinds of other agents, who we believe are morally responsible, so we should be consistent and adopt the non-historical compatibilist's stance on locally manipulated agents.

4.2. Five Unsuccessful Attempts to Respond to Manipulation Arguments

Many objections have been developed in the literature in response to global and local manipulation arguments. Though they can be adopted by non-historical compatibilists, I believe that they are inadequate, and should not be utilised in defending non-historical compatibilism.

4.2.1. *Intentional and Nonintentional Compulsion*

The first response I will consider seeks to find a relevant difference between a manipulated agent and a normally functioning agent. The most obvious difference between the two is that manipulation is unnatural; it artificially induces a set of pro-attitudes in a person. But if you are an ordinary agent living in a deterministic universe, you are not necessarily the subject of malicious brainwashing. Your pro-attitudes arise naturally and are not the effects of any vindictive neuroscientists.⁸⁶ The purpose of noting this distinction is to argue that compatibilists do not need to posit a historical condition in response to the local manipulation argument. Instead, they can just propose a clause concerning the genesis of an agent's pro-attitudes which states that one cannot be free or responsible when acting from an unnaturally caused desire.

Though that is a distinction, I do not think it is a relevant one. Mele notes that we could reimagine a case of manipulation so as to not include the nefarious intentional agents. For example, Beth could have come to have Chuckian values by passing through a strange electromagnetic field when flying over the Bermuda Triangle, which then leads her to

⁸⁶ A similar reply, but to a manipulation argument for incompatibilism, is explored and rejected in Kane, *Free Will and Values*, 38-46; *The Significance of Free Will*, 64-70.

murder Frank.⁸⁷ In these situations, no deliberate or purposeful agent caused Beth to have Chuckian values. Yet this new etiology is just as responsibility-nullifying as intentional manipulation; there is no relevant difference between the two types of manipulation. If you did not judge Beth as responsible before, you do not now have a reason to think she is responsible for killing her neighbour. When determining whether Beth is responsible or not, the relevant facts are Beth's value change and the manipulation. Whether the manipulation was intentional or nonintentional is extraneous to questions of Beth's responsibility.

Furthermore, even if this reply can be used to avoid the manipulation argument, it is not obvious that it will assist non-historical compatibilists. To deploy this response, one must assert that there is an important difference between desires that originated from natural causes, and desires that are introduced by brainwashers. But this seems to be a distinction based on the desires' histories. It is akin to Mele's negative historical compatibilism that states that you are unfree with respect to bypassed pro-attitudes. If one wishes to defend non-historical compatibilism, then this objection to the manipulation argument will not be of service as it commits one to a historical condition.

4.2.2. *Manipulated Beth is a New Person*

One obvious rejoinder to manipulation arguments is that Beth post-manipulation is a different person to Beth pre-manipulation. Beth's sudden revision in her values is so drastic that she bears little resemblance to her previous self. This gives a non-historical compatibilist an avenue to reply to the historical compatibilist's manipulation argument. If this contention about Beth's personal identity is correct, they can argue that our intuition is

⁸⁷ Mele, *Autonomous Agents: From Self-Control to Autonomy*, 168.

right in stating that pre-manipulation Beth is not responsible for her neighbour's murder, and that is because a new person, post-manipulation Beth, is the person responsible for the murder.

This objection is contingent on two views. The first is a theory of personal identity. As the only difference between pre- and post-manipulated Beth are her mental states, a non-historical compatibilist would have to hold that two time-slices of a human comprise a single person only when there is psychological continuity between the two time-slices.

'Psychological continuity' exists when there is a sufficient number of psychological connections, and 'psychological connections' refer to mental states such as memories.⁸⁸ A non-historical compatibilist who wishes to put forward this objection must deny other theories of personal identity, such as the view that bodily continuity, or the existence of a continued unified consciousness constitutes personal identity. This will be problematic for non-historical compatibilists who do not wish to endorse this metaphysical view. The second claim this objection rests on is that manipulation necessarily erases enough psychological connections to make post-manipulation Beth a new person. If the neuroscientists erased all remnants of Beth's personality, and implanted a completely new set of pseudo-memories and pro-attitudes, then it would be plausible to judge that there are two Beths, not just one continuous Beth.

However, even if we assume that psychological continuity is necessary and sufficient for personal identity, I believe that it is credible to suppose that manipulation will not necessarily interrupt psychological continuity between two time-slices of an agent, so we should not rest our objection on this point. Mele's manipulators can leave Beth's memories intact and only alter a small segment of her values. It is intuitive that Beth is still the same

⁸⁸ This is similar to the account given in Derek Parfit, *Reasons and Persons* (Oxford: Oxford University Press, 1984), 204-09.

person if she undergoes manipulation of this nature. This can be seen if we imagine Beth's thought processes when she wakes up. Beth could reflect on what she has done the previous day, and may take time to think of her family, her friends, and the first school she attended. Beth also remembers being a good person and is surprised to discover she no longer cares about leading a moral life and instead strongly desires to commit murders.

I find the above story makes it intuitive that Beth is still the same person because her memories psychologically link her to a time-slice prior to the manipulation. Moreover, manipulated Beth has more psychological continuity with pre-manipulated Beth than an ordinary person has with their infant self, and if we are willing to treat the latter as one entity then we should say that personal identity has been preserved after manipulation.⁸⁹ Although I reject this particular argument from personal identity, I believe a similar point can be made, in non-historical compatibilism's favour, without assuming that Beth's identity has been compromised.⁹⁰ I will address this point in Section 4.11.

4.2.3. *Chuck as Responsible for More than Beth*

Non-historical compatibilists acknowledge that it is intuitive that there is some difference between Chuck and Beth. McKenna believes that this intuition is not contrary to non-historical compatibilism because we can point out a difference between Chuck and Beth without sacrificing our ahistoricism. McKenna does this by arguing that the difference between the two is that though both are responsible for murder, only Chuck is responsible

⁸⁹ Mele discusses the objection from personal identity in Mele, *Autonomous Agents: From Self-Control to Autonomy*, 165, footnote 22.

⁹⁰ Benjamin Matheson has argued that though historical compatibilists can insist that Beth is still the same person after being manipulated, this has counterintuitive ramifications. See Matheson, "Compatibilism and Personal Identity."

for becoming the kind of person who would murder. Beth, on the other hand, is only responsible for the murder because she did not choose to have the character of a cold-blooded killer. McKenna believes that this difference explains our intuition about Chuck and Beth, and that this difference is something non-historical compatibilists can happily acknowledge.⁹¹

This argument might be acceptable if the intuition from Mele's manipulation argument was that there is *some* difference between Chuck and Beth, where we had not quite put our finger on what that difference might be. But the intuition is much more substantial than that. It is not merely that there is some difference, but that there is a fundamental difference; Chuck is responsible for the murder he performs and Beth is not. It will be of no help to tell someone who is agnostic about the issue that Chuck is responsible for some extra things while Beth is only responsible for one action, especially when the agnostic is already inclined towards scepticism about whether Beth is responsible at all.

4.2.4. *Manipulation as a Moral Violation*

Nomy Arpaly's method for distinguishing between the cases, is to argue that Beth has had a moral wrong done to her, whereas Chuck has not. Beth's autonomy is violated by the brainwashing. Beth had her own set of values which she had been developing over her lifetime, and suddenly this fundamental aspect of her was snatched away by

⁹¹ McKenna, "Responsibility and Globally Manipulated Agents," 183; "Defending Nonhistorical Compatibilism: A Reply to Haji and Cuypers," 271; "Moral Responsibility, Manipulation Arguments, and History: Assessing the Resilience of Nonhistorical Compatibilism," 163.

neuroscientists, which is morally wrong. Chuck's case does not share these features, which is why, according to some non-historical compatibilists, we still believe he is blameworthy.⁹²

This observation can be used to defend non-historical compatibilism in one of two ways. McKenna believes that the moral wrong done to Beth may partly explain what gives rise to our intuition in Mele's example. Beth's being wronged makes us hesitant to blame her. Though she is in fact responsible for what she does, we know she has suffered a substantial loss and are unsure whether she deserves blame given the kind of past she has. Knowledge of these facts suppresses the intuition that Beth is responsible for the murder. McKenna counsels us by demonstrating that we can separate the fact that Beth has been treated poorly from the fact that she is a responsible agent.⁹³ We can accept the non-historical conclusion about this case whilst experiencing empathy for Beth.

A second way of utilising this response is to insist that there is a relevant difference between manipulated agents and non-manipulated agents. Manipulated agents have had their autonomy violated while normal agents have not been wronged. Compatibilists who accept this version of the reply will then be faced with the task of developing a compatibilist account that explains and analyses how violations of autonomy disrupt agency.

As disturbing as they may be, being morally wronged is not the kind of difference we are looking for. To demonstrate this, consider cases of non-intentional manipulation (from Section 4.2.1). In cases of non-intentional manipulation, there are no nasty neuroscientists intending to do Beth harm. Instead, Beth merely flies past the Bermuda Triangle. Though there is no one with hostile intentions doing harm to Beth, we are still inclined to treat the non-intentional manipulation cases and the intentional manipulation cases symmetrically; if

⁹² Nomy Arpaly, *Unprincipled Virtue* (New York: Oxford University Press, 2003), 128.

⁹³ McKenna, "Responsibility and Globally Manipulated Agents," 183-84; "Defending Nonhistorical Compatibilism: A Reply to Haji and Cuypers," 271-72; "Moral Responsibility, Manipulation Arguments, and History: Assessing the Resilience of Nonhistorical Compatibilism," 163-64.

one is responsibility-undermining, then both are responsibility-undermining. In cases where the immoral nature of the brainwashing has been subtracted, we still believe Beth is not responsible. Therefore, it is not the immoral nature of the brainwashing that effects our judgements of Beth and neither version of the response succeeds.

4.2.5. *Harrison's Attempt to Transform Historical Compatibilists Into Non-historical Compatibilists*

Gerald K. Harrison has provided a novel argument for turning historical compatibilists into non-historical compatibilists.⁹⁴ He believes that as compatibilists do not believe there is a relevant difference between compatibilist agents and libertarian agents, then if it can be shown that manipulated libertarian agents are morally responsible, then we must think the same of manipulated compatibilist agents.

Take the case of Beth, but recast it so that Beth not only possesses compatibilist control, but also has agent-causal libertarian control over her actions. Libertarian Beth is the ultimate source of her actions, the uncaused cause of her behaviour, and has legitimate alternative possibilities open to her when she acts. Imagine that Beth has been approached by her wealthy brother Frank for help, and Beth is torn. As neuroscientists have recently manipulated her to make her greedy, she wishes to murder him, but she also has a strong sense of familial obligations and also desires to assist Frank. Beth's conflicting motives mean that at the time of the decision, Beth has reasons for performing either action open to her,

⁹⁴ Gerald K. Harrison, "A Challenge for Soft Line Replies to Manipulation Cases," *Philosophia* 38, no. 3 (2010).

and her character does not determine which of these outcomes will be brought about.

Whichever way Beth decides, the decision will reinforce that aspect of her.⁹⁵

Suppose further that, as in the original case, Beth was a compassionate person and only has the character she does because she has recently been manipulated. The neuroscientists want Beth to have Frank's money, and do not mind which action she performs because either way she will get it (if Beth helps Frank, he will be so grateful that he will donate his fortune to her there and then). Harrison believes that manipulated Beth with libertarian control is responsible to some degree for whatever action she performs. Though Beth did not have control over the assortment of options before her, she does have control over which of the options is actualised, so she is morally responsible for either murdering or supporting Frank.

To demonstrate there is no difference between libertarian agents and compatibilist agents, we need to subtract the libertarian elements of the story. Reimagine the case but remove Beth's ability to agent-cause her actions. Because compatibilists do not think ontologically irreducible agent-causation is necessary for moral responsibility, this change in the case should not influence our intuitions. We should still think that Beth, with her ability to do otherwise, is morally responsible. Then, present the case where Beth no longer has alternative possibilities. She may still struggle with what action to undertake, but the decision's upshot is determined. This is a standard case of exercising compatibilist agency so, to be consistent, we should believe manipulated compatibilist Beth is morally responsible.

Because we are compatibilists, we do not think that it is the libertarian features of the previous story that explain why Beth was partly morally responsible for what she does. This

⁹⁵ Clarke, *Libertarian Accounts of Free Will*.

is due to the fact that compatibilists, including myself, tend to avow that if there is a coherent picture of libertarian agency, then it is no better at grounding moral responsibility than its closest compatibilist cousin. So, given our compatibilist outlook, we cannot say that there is any difference between the previous cases of libertarian Beth and compatibilist Beth.

Now there is a problem on the historical compatibilists' hands. We judge that libertarian Beth is morally responsible though she was manipulated, and we believe that libertarianism is no better a theory than compatibilism. So we must deem that manipulated libertarian Beth is no different to manipulated compatibilist Beth and that both are morally responsible. This is how Harrison believes compatibilists can become committed to non-historical compatibilism.

However, I believe that the argument fails. I agree that libertarian Beth is at least partly responsible for what she does, but that is because I already do not view manipulation as a threat to compatibilist Beth. But for the historical compatibilists who do perceive manipulation as vicious, I believe that they will perceive libertarian Beth as not being morally responsible because of her history of manipulation.⁹⁶ This would not be a surprising reaction. After all, the decision made available to Beth has been provided entirely by her brainwashers. If one possible outcome determined by neuroscientists annuls responsibility, then it would be of no help to add a second possible outcome that was also administered by neuroscientists. Both libertarian Beth and compatibilist Beth invalidate the historical compatibilists' historical condition, so she will not evaluate either as morally responsible.

⁹⁶ For authors who have put forward soft compatibilist theories and view libertarianism as susceptible to manipulation, see Ishtiyaque Haji and Stefaan E. Cuypers, "Libertarian Free Will and CNC Manipulation," *Dialectica* 55, no. 3 (2001).

Harrison anticipates this historical compatibilist's objection and gives two responses to the idea that agents need to have a history in order to be morally responsible. The first response is that it is intuitive to think that, in situations where the manipulation has left Beth's libertarian free will undamaged, she is still morally responsible. This amounts to merely a denial of my previous claim that historical compatibilists do not find it intuitive that Beth is morally responsible. Without further argument, this response leads only to a stalemate.

Harrison's second response is to point to 'magical agents' and argue that they are similar to manipulated libertarian agents.⁹⁷ One kind of magical agent is Mele's minutelings. Harrison believes that a minuteling clone of Beth, with libertarian powers, is a being that is morally responsible for what she does. Minuteling libertarian Beth does not have a history, but this does not remove her moral responsibility. Compare her to manipulated libertarian Beth, who has been brainwashed so that she is the psychological doppelganger of minuteling libertarian Beth. Given that minuteling libertarian Beth's lack of history does not rob her of her moral responsibility, Harrison believes that we should similarly not take into account manipulated libertarian Beth's history and instead grant the fact that she is morally responsible for what she does.

Though I agree that one should give symmetrical judgements to minuteling libertarian Beth and manipulated libertarian Beth, I do not think Harrison has provided historical compatibilists with a reason to accept the symmetry thesis. Negative historical compatibilists will reject Harrison's reasoning behind his symmetrical judgement, because they believe that if agents do have a history, it must be a history of a certain kind. Therefore, they can accept that minutelings are responsible for what they do, but deny that

⁹⁷ The term 'magical agent' comes from "Magical Agents, Global Induction, and the Internalism/Externalism Debate."

ordinary humans who have been manipulated are responsible. Moreover, historical compatibilists have criticised this symmetry thesis.⁹⁸ So Harrison must provide a stronger defence of the symmetry thesis if he wishes to make his original argument convincing.

Harrison requires an argument in defence of the symmetry thesis to support his argument, so that historical compatibilists have a reason to give up their historical condition. But a successful argument in defence of the symmetry thesis would be sufficient on its own for refuting negative historical compatibilism, as it would show that it is untenable to think that there is a relevant difference between magical agents and manipulated agents. Harrison's argument comparing manipulated libertarian Beth with manipulated compatibilist Beth then begs the question, because one must accept the symmetry thesis for it to be convincing, yet anyone who accepts the symmetry thesis will already have accepted non-historical compatibilism. Therefore, we should shift our attention away from Harrison's argument as it is a moot point, and instead focus on reasons why we should accept the symmetry claim. I will take up this task in Section 4.11.

4.3. Manipulated Histories

I will now focus on McKenna's inventive argument that undermines the motivation for maintaining historical compatibilism. He has demonstrated that though what causes compatibilists to hold a historical condition is a desire to elude manipulation arguments, historical compatibilists are only able to circumvent *local* manipulation arguments, but must bite the bullet on *global* manipulation arguments. I will present his reasoning and develop

⁹⁸ Mele, "Moral Responsibility and History Revisited," 466.

his point further by arguing that there is no basis for historical compatibilists to assert that a significant distinction can be made between the two kinds of manipulation.

Historical compatibilists posit historical conditions as they believe that this will assist them in evading manipulation arguments. But as McKenna has observed, this can only forestall the inevitable; though it may block the local manipulation argument, a global manipulation argument can be devised whose conclusion historical compatibilists are forced to accept. This is because historical compatibilists are compatibilists. So long as their historical conditions are compatible with naturally occurring determinism, then it is conceptually possible that their historical conditions are compatible with a world that is intentionally designed by an evil demon, an artificial analogue of determinism.⁹⁹ Since historical compatibilism can be hoisted by its own petard, McKenna argues that it should be abandoned as it cannot combat manipulation arguments in general.¹⁰⁰

To illustrate this point more clearly, I will discuss Derk Pereboom's global manipulation argument, dubbed the Four-Case Argument. Originally designed to defend incompatibilism and attack compatibilism, I will retool it so that it poses a specific kind of problem for historical compatibilism. Pereboom begins his argument with a case of what is supposed to be responsibility-undermining manipulation, and presents a series of slightly tweaked cases until he arrives at a case of regular determinism. In each case, Professor Plum murders Ms. White for egoistic reasons, while satisfying many compatibilist control conditions of moral responsibility plus historical conditions, yet, supposedly, we do not believe Plum is blameworthy for killing White in any of the cases. As there is no obvious difference between the first and final case, the argument is intended to be a *reductio ad absurdum* against

⁹⁹ Watson, "Some Worries About Semi-Compatibilism: Remarks on John Fischer's the Metaphysics of Free Will," 141.

¹⁰⁰ McKenna, "A Hard-Line Reply to Pereboom's Four-Case Manipulation Argument," 143-44.

compatibilism because it seems to compel compatibilists to say that manipulated agents are morally responsible.

In Case 1, neuroscientists produce in Plum a 'neural state' that inclines the way he behaves. Producing a neural state in someone is supposed to be akin to, for example, a person being insulted without reason, causing them to be angry. This neural state of anger influences their behaviour, so that they act angrily and less charitably for the rest of the day. In these cases, the anger is induced by external factors, yet we do not think the person is no longer an agent or excused for acting rudely or poorly. Plum's neural state in Case 1 is comparable to the state of anger.¹⁰¹

Case 1: A team of neuroscientists has the ability to manipulate Plum's neural states at any time by radio-like technology. In this particular case, they do so by pressing a button just before he begins to reason about his situation, which they know will produce in him a neural state that realises a strongly egoistic reasoning process, which the neuroscientists know will deterministically result in his decision to kill White. Plum would not have killed White had the neuroscientists not intervened, since his reasoning would then not have been sufficiently egoistic to produce this decision. But at the same time, Plum's effective first-order desire to kill White conforms to his second-order desires. In addition, his process of deliberation from which the decision results is reasons-responsive; in particular, this type of process would have resulted in Plum's refraining from deciding to kill White in certain situations in which his reasons were different. His reasoning is consistent with his character because it is frequently egoistic and sometimes strongly so. Still, it is not in general exclusively egoistic, because he sometimes successfully regulates his behaviour by moral reasons, especially when the egoistic reasons are relatively weak. Plum is also not constrained to act as he does, for he does not act because of an irresistible desire – the neuroscientists do not induce a desire of this sort.

Plum fulfils a plethora of compatibilist control conditions, yet Pereboom does not believe that Plum is morally responsible for what he does in this case. Pereboom is aware that compatibilists might reply to Case 1 by arguing that it is the fact that Plum's brain can be

¹⁰¹ The updated version of Case 1 was inspired by Seth Shabo, "Uncompromising Source Incompatibilism," *ibid.* 80, no. 2 (2010): 376. It was originally presented in Pereboom, *Free Will, Agency, and Meaning in Life*, 76-77. The remainder of the Four-Case argument was first presented in "Determinism *Al Dente*," *Living without Free Will*, 100-26.

directly influenced at any given moment that drives our intuitions and not the deterministic manipulation per se. But Pereboom addresses this with Case 2:

Case 2: Plum is like an ordinary human being, except that he was created by neuroscientists, who, although they cannot control him directly, have programmed him to weigh reasons for action so that he is often but not exclusively rationally egoistic, with the result that in the circumstances in which he now finds himself, he is causally determined to undertake the moderately reasons-responsive process and to possess the set of first- and second-order desires that results in his killing Ms. White. He has the general ability to regulate his behaviour by moral reasons, but in these circumstances, the egoistic reasons are very powerful, and accordingly he is causally determined to kill for these reasons. Nevertheless, he does not act because of an irresistible desire.

Plum is now no longer manipulated from moment to moment. Instead, he is manipulated at his birth. Pereboom says the fact that Plum is manipulated by factors beyond his control explains why we do not think Plum is responsible in Case 1 or 2. Pereboom then argues that these same responsibility-undermining features are present in Case 3, which bears a stronger resemblance to a real-world scenario:

Case 3: Plum is an ordinary human being, except that he was determined by the rigorous training practices of his home and community so that he is often but not exclusively rationally egoistic (exactly as egoistic as in Cases 1 and 2). His training took place at too early an age for him to have had the ability to prevent or alter the practices that determined his character. In his current circumstances, Plum is thereby caused to undertake the moderately reasons-responsive process and to possess the first- and second-order desires that result in his killing White. He has the general ability to grasp, apply, and regulate his behaviour by moral reasons, but in these circumstances, the egoistic reasons are very powerful, and hence the rigorous training practices of his upbringing deterministically result in his act of murder. Nevertheless, he does not act because of an irresistible desire.

Pereboom states that Plum's non-responsibility in Cases 1 and 2 generalises to Case 3 as they all involve Plum killing Ms. White due to factors outside of Plum's control. And if Case 3 is responsibility-undermining, then the ordinary Case 4 will be as well:

Case 4: Physicalist determinism is true, and Plum is an ordinary human being, generated and raised under normal circumstances, who is often but not exclusively rationally egoistic (exactly as egoistic as in

Cases 1-3). Plum's killing of White comes about as a result of his undertaking the moderately reasons-responsive process of deliberation, he exhibits the specified organisation of first- and second-order desires, and he does not act because of an irresistible desire. He has the general ability to grasp, apply, and regulate his behaviour by moral reasons, but in these circumstances the egoistic reasons are very powerful, and together with background circumstances they deterministically result in his act of murder.

Pereboom believes that there is no relevant difference between Cases 3 and 4, and so the non-responsibility of Plum generalises to Case 4. This is a challenge for compatibilism because it seems that Cases 1-4 are similar in respects that would bear on free will and moral responsibility, yet Case 1 is a case of global manipulation and Case 4 is an instance of ordinary determinism.

Given that their historical conditions are compatible with ordinary determinism, historical compatibilism now faces a counterexample because their historical conditions are also compatible with global manipulation. Fischer & Ravizza have attempted to proffer a historical condition that rules out global manipulation. Fischer & Ravizza require that an agent see herself as able to affect the world in various ways, and understand that she is a fair target of the reactive attitudes.¹⁰² Since Plum has been manipulated prenatally, his life from birth could have been identical to yours or mine, and there is no reason to suppose that he would not have the beliefs that Fischer & Ravizza cite above. In anticipation to this objection, Fischer & Ravizza stated that:

"[T]he agent's view of himself must be based on his evidence in an appropriate way. (...) This condition is intended (in part) to imply that an individual who has been electronically induced to have the relevant view of himself (and thus satisfy the first two conditions on taking responsibility) has *not* formed his view of himself in the relevant way."¹⁰³

But it is puzzling how this condition is supposed to achieve excluding Plum from responsibility. The idea, I gather, is that Plum is unaware of the causal origins of his

¹⁰² Fischer and Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, 210-14.

¹⁰³ *Ibid.*, 236.

motivations, and so his beliefs are not justified by his evidence. However, this is not just true of Plum, but of nearly all human beings throughout time. Most persons are unaware of the details of the neurology that underlie their mechanisms of action. But Fischer & Ravizza do not believe that this renders us non-responsible for what we do. In fact, Fischer & Ravizza argue that we do not need to know the details of our neurology in order to take responsibility for our mechanisms.¹⁰⁴ Since ordinary agents can take responsibility and have their beliefs based appropriately on the evidence despite being unaware of the causal origins of their behaviour, Fischer & Ravizza cannot deny Plum responsibility for that very same reason.¹⁰⁵

Following the advent of the Four-Case Argument, Fischer has changed tack in his response to globally manipulated agents. He now argues that Plum is morally responsible for the murder, but is not blameworthy.¹⁰⁶ I will not elaborate on Fischer's distinction between responsibility and blameworthiness. What is important is that despite being a positive historical compatibilist, he agrees with non-historical compatibilists on the subject of global manipulation. This is an important concession, because I believe it helps undermine our reasons for accepting historical compatibilism.

Mele, the negative historical compatibilist, also holds that globally manipulated agents are morally responsible. Mele has generated his own global manipulation case, called the *Zygote Argument*. Suppose that a deity Diana creates a zygote in Mary. Diana deduces from the deterministic laws and a complete description of the universe at the time that this zygote will mature into an ideally self-controlled reliable deliberator named Ernie. Diana

¹⁰⁴ Ibid., 233-35.

¹⁰⁵ Pereboom, *Living without Free Will*, 120-22; Kapitan, "Autonomy and Manipulated Freedom," 90-94.

¹⁰⁶ John Martin Fischer, "Responsibility and Manipulation," *The Journal of Ethics* 8, no. 2 (2003): 154-59; *My Way: Essays on Moral Responsibility* (New York: Oxford University, 2006), 230-34; "Response to Kane, Pereboom, and Vargas," in *Four Views on Free Will*, ed. John Martin Fischer, et al. (Malden, Mass.: Blackwell Publishers, 2007), 185-88.

also desires that an event *E* occurs in 30 years, and knows that by manufacturing Ernie's zygote the way she did, *E* will be brought about. Diana is criminally insane and cannot comprehend morality. Mele then argues:

- 4.1. Because of the way his zygote was produced in his deterministic universe, Ernie is not a free agent and is not morally responsible for anything.
- 4.2. Concerning free action and moral responsibility of the beings into whom the zygotes develop, there is no significant difference between the way Ernie's zygote comes to exist and the way any normal human zygote comes to exist in a deterministic universe.
- 4.3. So determinism precludes free action and moral responsibility.¹⁰⁷

We do not need to defend or attack any of the premises here. All that needs to be shown is that Mele believes compatibilists should deny premise 1, saying "[g]iven their compatibilism and the assumption that premise 2 is true, Ernie should strike them as free and morally responsible (in light of his properties as an agent)."¹⁰⁸ Despite denying that manipulated Beth is responsible for what she does, Mele believes that compatibilists should accept that globally manipulated Ernie is morally responsible.

Pereboom's Four-Case Argument and Mele's Zygote Argument have illustrated that the major historical compatibilists in the literature, who defend two different varieties of historicism, are forced to accept that global manipulation does not result in non-responsible agents. If we make slight adjustments to the manipulation, namely that we have it occur

¹⁰⁷ Mele discusses the Zygote Argument in Mele, *Free Will and Luck*, 184-95; "Manipulation, Compatibilism, and Moral Responsibility," 278-84; "Moral Responsibility and Agents' Histories," 179, footnote 24; "Manipulation, Moral Responsibility, and Bullet Biting," 175-83.

¹⁰⁸ *Free Will and Luck*, 193.

before the agent's birth or have it orchestrated by an omniscient god, then historical compatibilists are compelled to alter their judgements about the situation.

Non-historical compatibilists are not immune to the global manipulation argument; their conditions of moral responsibility are also compatible with global manipulation. But I believe that global manipulation is particularly problematic only for historical compatibilism. This is because non-historical compatibilists already contend that locally manipulated agents are free and responsible, so it does not seem especially burdensome for them to hold the same of globally manipulated agents.

Additionally, they must defend their ahistoricism, so non-historical compatibilists will already have arguments that justify their position that locally manipulated agents are responsible. I believe that most of these arguments can be co-opted and used to defend the non-historical compatibilist's stance on globally manipulated agents. Thus, whatever strategy that is used for dealing with local manipulation arguments can be applied to global manipulation arguments. But historical compatibilism is committed to the view that the arguments against local manipulation fail and so cannot rely on them to ward off the threat of global manipulation. Non-historical compatibilism already has some tools for dealing with manipulation, whilst historical compatibilism has none.

Historical compatibilists believe that local manipulation arguments successfully demonstrate the flaws in their opponent's position. But as they are compatibilists, they must believe that global manipulation arguments fail. I find this a puzzling position, because, as stated earlier, it seems that the main motivation behind accepting a historical condition is trying to evade manipulation arguments. But now historical compatibilists must agree that their historical condition is of no use when it comes to globally manipulated agents, due to

their compatibilist leanings.¹⁰⁹ And so the main support for historical compatibilism is undermined.

Historical compatibilists can reply by arguing that it is intuitive that only local manipulation eliminates responsibility and that there is not a similar intuition behind global manipulation cases. If true, this would mean that they are not obliged to find a response to global manipulation arguments, as their theory is already protected from the only intuitively threatening kind of manipulation. However, I believe this is unconvincing for two reasons.

Firstly, it seems strikingly obvious that Pereboom's Four-Case Argument and Mele's Zygote Argument are powerful in virtue of the fact that they enshrine pre-theoretical commitments we share. We would not find global manipulation arguments persuasive if there was no intuition backing them. Secondly, even if it is true that large subsets of the population do not share any global manipulation intuitions, they would also not have any local manipulation intuitions either. That is, I expect that individuals who find one version of manipulation troublesome will feel the same about the other. Historical compatibilism's core audience will feel the tug of the global manipulation arguments. Individuals who do not have the global manipulation intuition will also not have the local manipulation intuition, and will most likely be non-historical compatibilists or free will sceptics. So historical compatibilists cannot claim that their position captures all our intuitions on these issues.

Given that historical compatibilists hold differing stances on local and global manipulation, they must assert that there is a relevant distinction between the two with respect to free will and moral responsibility. Historical compatibilists are committed to holding that there are properties of local manipulation that render it a responsibility-

¹⁰⁹ This may appear as if I am suggesting that incompatibilists are immune to manipulation arguments, though I do not endorse that claim. For an argument that multiple kinds of incompatibilist positions can fall prey to manipulation cases, see Haji and Cuypers, "Libertarian Free Will and CNC Manipulation."

nullifying process, but these properties are not shared by global manipulation. Though they may be unable to pinpoint what exactly are the properties that make this substantial difference, historical compatibilists can still aver that they exist, though they are yet to be found.

Though this could be the case, I believe that this is *prima facie* improbable. This is because this reply hinges on their being a substantial difference between manipulation that interrupts the natural progression of one's life, and manipulation that begins early in one's life. The manipulation in both kinds of cases is too similar for us to expect that they will have different explanations for why they nullify moral responsibility. Both cases feature intentional manipulation of an agent's values so that they will act in specific ways. The difference in when the manipulation occurs is not significant enough to warrant two independent explanations of how manipulation undermines responsibility. Any motivation one has to accept that Beth is not responsible should also generalise to Plum. However, historical compatibilists must reject these claims, though the grounds for doing so are mysterious.

Though we can direct these objections from global manipulation at historical compatibilism, it does not necessarily make it inferior to non-historical compatibilism. Historical compatibilists may insist that, although they must counterintuitively believe that globally manipulated agents are free and responsible, their position is still overall preferable to non-historical compatibilism. This is because both parties to the debate bite the bullet on global manipulation. But on issues of local manipulation cases, only historical compatibilism delivers the intuitive result. Non-historical compatibilism states that locally manipulated Beth is responsible, a claim that is *prima facie* absurd. Historical compatibilists can rebuff

the current argument by noting that their view is still overall more intuitive and therefore preferable to non-historical compatibilism.

Although this is correct, I am uncertain that this is a credit to historical compatibilism. It renders historical compatibilism *ad hoc* in a troubling manner. There is, from what I can see, no principled reason to distinguish between manipulation that began before the actor was born and manipulation that interrupts the actor's life. It seems odd to claim that lifelong manipulation (that is, global manipulation) is less pernicious to moral responsibility than the one-off local manipulation in a person's life.

Furthermore, it is unclear why historical compatibilists are willing to believe that globally manipulated agents are free, yet deny that locally manipulated agents are free. I do not see the reason for treating the two cases asymmetrically. If I were to hazard a guess, I would say the motivation for these divergent stances is that historical compatibilists believe the intuitions behind compatibilism are weightier than the intuitions behind global manipulation, whereas they find the intuitions for local manipulation stronger than the intuitions for ahistoricism. Attempting to capture intuitions in this way causes trouble for historical compatibilism when we notice just how similar global and local manipulation are.

Historical compatibilists can still believe they maintain the advantage in this specific debate because they think their position is more intuitive overall than non-historical compatibilism. They assume that it is intuitive that agents like locally manipulated Beth are not morally responsible. But I reject this assumption. After having built up manipulation arguments so that they appeared to undermine responsibility, I will proceed to weaken the

arguments and demonstrate that one can be manipulated yet still remain morally responsible.¹¹⁰

4.4. Intuitions about Manipulated Agents Performing a Good Action

In the Four-Case Argument, Pereboom's Case 2 is designed to elicit incompatibilist intuitions. Dana Kay Nelkin has modified the description of the case so that it gives rise to intuitions that support the view that Plum has free will. By altering the case to one where Plum performs a good deed, Nelkin has made it persuasive that a manipulated Plum is still morally responsible. If it can be shown that it is intuitive that globally manipulated agents are praiseworthy for what they do, then this weakens the support the Four-Case Argument can provide incompatibilism. I believe Nelkin's counterexample can be adapted into a similar objection against Mele's local manipulation argument, so that it undermines its ability to support historical compatibilism. The edited case is as follows:

A team of neuroscientists kidnap Chuck and alter his neurology so that he fully understands and recognizes good reasons for acting. He is moved by people in distress and desires to help them so as to relieve their suffering. Suppose that one day he finds himself in a situation in which he can help a child only at great risk to himself. He thinks about the relevant considerations and decides that all things considered, helping is the right thing to do and resolves to do it.¹¹¹

In this example, it is intuitive that Chuck is morally responsible for what he does. Our intuitions about local manipulation cases are now in conflict. We believe that manipulated Beth who performs a bad act is not responsible, whereas we believe Chuck who performs a

¹¹⁰ This is noted but not considered a problem for historical compatibilism in McKenna, "Responsibility and Globally Manipulated Agents," 176-79.

¹¹¹ This case is inspired by Nelkin, *Making Sense of Freedom and Responsibility*, 56-57.

good act deserves credit. Moreover, it is not clear which recalcitrant intuition we should abandon.

Pereboom has responded to the global manipulation version of this case, arguing that the intuition from this case does not support the view that Plum is praiseworthy in a deserving sense. He notes that we ordinarily praise persons for qualities that they have done nothing to acquire. For example, we praise a gifted student who does well on an exam without studying, or celebrate the success of an impressive athlete. In both cases we praise, despite the fact that these achievements are largely due to factors beyond their control. As we frequently praise individuals who do not fundamentally deserve praise, Pereboom believes that the fact that we want to praise Chuck is not evidence that Chuck deserves praise.¹¹²

Although it is true that we occasionally praise agents when they are not praiseworthy, I do not believe that this objection can be used successfully by a historical compatibilist to defend against Nelkin's case. I believe this response commits you to an incompatibilist theory of freedom and responsibility, which is why it can be meted out by Pereboom but not Mele. The reply is incompatibilist in nature as it is trying to draw a link between good Chuck and the impressive athlete. Though we praise the athlete, she is not praiseworthy because her actions are the outcome of her genetics for which she is not responsible. If we are to apply this to Chuck, we would have to say that though we are praising Chuck, this is not evidence that he is praiseworthy because his action is just the expression of a character that he is not responsible for, but is in fact the work of neuroscientists.

The problem with taking such a stance is that regular individuals are no different to Chuck in this regard. Our characters do not ultimately come from exercising our agency, but

¹¹² Pereboom, *Free Will, Agency, and Meaning in Life*, 102.

are the results of our genes, environment, and upbringing.¹¹³ If we accept Pereboom's reason for denying that Chuck is praiseworthy, then we must also say that no one is responsible for their actions as they are all the results of traits and talents that are undeserved. Historical compatibilists should reject this reason as it leads to incompatibilism. Pereboom is correct in stating that when we praise an agent, they are not necessarily praiseworthy. However, his reasoning for distinguishing the two concepts in this situation cannot be successfully endorsed by compatibilists. I will continue to develop my defence of this variant of Nelkin's argument in Sections 4.5 and 4.6.

4.5. Manipulation and the Consequentialist Aspects of Praise and Blame

I believe that we have good normative reasons to accept that locally manipulated agents are free and responsible, even if we do not possess that intuition. If we accept that our practice of praising and blaming others enshrines consequentialist attitudes, then that gives us reason to believe locally manipulated agents are free and responsible because our reactive attitudes are able to influence their future behaviour as much as it can shape an ordinary individual's future actions. This argument is distinct from the others in this chapter in that it is arguing for a normative claim, that we *should* adopt non-historical compatibilism, and does not rely on my reader actually finding it intuitive that non-historical compatibilism is true.

When we normally discuss the concepts of praise and blame, we are interested in who deserves praise or blame based on their past behaviour. But the conditions for when a

¹¹³ Compatibilists freely admit that we can change our character by exercising our agency. But that does not mean that our character *ultimately* comes from our agency. The very first character we have is not something we have control over.

person is praiseworthy or blameworthy can come apart from the conditions for when we should praise or blame that person. Consider Irena Sendler, who smuggled Jewish children out of Warsaw when it was occupied by Nazi Germany. She is praiseworthy for performing such a noble deed with great risk to herself. However, it might be unwise to praise her because this could draw the attention of the Nazis, who would then injure her. Similarly, it is reasonable for me to hesitate blaming you for treading on my foot if I know this will cause you to receive a punishment that is not a suitable response to your action.

Questions of whether someone deserves praise or blame are distinct from, though related to, questions of whether we ought to, all things considered, praise or blame someone. Whether it is appropriate to offer credit or censure depends partially on whether doing so will lead to more good consequences than bad. Praising and blaming serve important social functions. When we praise someone, we reward them with the hope that it will encourage them or others witnessing the praising to perform a morally good action in the future. Likewise, when we blame and punish individuals, we are also trying to deter crime from occurring. These are beneficial, future-looking considerations that influence whether we should compliment or condemn someone's actions that are independent of whether the person deserves our reactions to them.

Our attitudes towards young children, animals and criminal policy reflect the consequentialist aspects of our praising and blaming practices. Young children and animals do not have the requisite cognitive capacities to be considered morally responsible for the actions they perform. But this does not inhibit us from praising or blaming them for the actions they perform. In these cases, we hope to direct their future behaviour by incentivising them to perform the good actions again, and making it costly to act harmfully. The fact that they do not deserve the praise or blame does not make it inappropriate to

praise or blame them. Similarly, many philosophers believe that capital punishment's potential to deter future offenders is a good reason to implement it for those who commit serious crimes.¹¹⁴ Focusing on the possible future goods gained from presently blaming someone demonstrates that there is also a consequentialist consideration to blame, which is independent of whether the person deserves the blame.

The consequentialist aspects of praising and blaming give us a normative reason to endorse non-historical compatibilism. If you accept that we care about the concept of moral responsibility and think it is important because we use the tools of blaming and praising to help correct behaviour, then you should also accept that manipulated agents can be praiseworthy or blameworthy for what they do. Manipulated agents are *ex hypothesi* just as able to reflect on what they do as an ordinary agent. They are still influenced by the way society reacts to and treats them, despite the manipulation altering who they once were. The social functioning element of moral responsibility will work successfully and can still be applied to manipulated agents.

One might object to this argument by noting that I am advocating a revision of our praising and blaming practices, using the example of manipulated agents. But as manipulated agents do not exist, the normative reason to endorse non-historical compatibilism is weak and we thus do not have any cause to change our practices. I believe the normative reason is still powerful as there are potentially a number of real life individuals that historical compatibilists will not consider morally responsible that consequentially we should still praise or blame. Consider Robert Harris from Chapter 1. It is plausible that some historical compatibilists will argue that he is not responsible for what he

¹¹⁴ Philosophical defences of the death penalty can be found in Ernest Van Den Haag, "On Deterrence and the Death Penalty," *The Journal of Criminal Law, Criminology and Police Science* 60, no. 2 (1969); David A. Conway, "Capital Punishment and Deterrence: Some Considerations in Dialogue Form," *Philosophy and Public Affairs* 3, no. 4 (1974).

did, given his history. However, it is also possible that he is responsive to being blamed by his community. This would then be another example where we have a normative reason to accept non-historical compatibilism and treat an agent as morally responsible because they could be educated by praise/blame and realise the consequentialist benefits of non-historical compatibilism.

This response to manipulation arguments is unique, as it does not dispute that it is intuitive that locally manipulated agents are morally responsible. Instead, it admits that we may have this intuition, but that we have better reasons to accept non-historical compatibilism. If we believe this is an important dimension to moral responsibility, then we should hold that manipulated agents are morally responsible for their actions.

4.6. Fixating on Agential Properties

One factor that partially explains the non-historical intuition from Nelkin's case of Chuck performing a good action is that the description of that case draws more attention to the agential abilities of Chuck. The details of the case include both the fact that Chuck is sensitive to reasons and that he has a complex set of emotional reactions to unfortunate events in the world. I believe that the most effective line of attack that the non-historical compatibilist can adopt is to highlight the many properties that a manipulated agent still has post-manipulation. Focusing on these aspects of the manipulated agent make us see that they are very much like you or me, that their actions can still be the result of a process of reflection, and that their minds do not function like an automaton or computer program. These sentiments are echoed by McKenna:

"The compatibilist's best strategy, it seems to me, is not to show how a suitably determined agent differs so very much from a globally manipulated agent. It is rather to show how similar they are. The compatibilist needs to make clear that once the manipulation is so qualified that all an agent's current time slice compatibilist-friendly structures are properly installed through a process of manipulation, then the role of the manipulator begins to shrink into the background; we are simply left with a normal person who happened to be brought into existence in a very peculiar manner."¹¹⁵

McKenna is referring to Pereboom's global manipulation argument but this tactic can be appropriated to combat the historical compatibilist's local manipulation argument. When we really bring into focus the "victim's" agential powers, it becomes intuitive that they are morally responsible for what they do, and the fact that they were manipulated begins to lose its grip. To demonstrate this, I will again present Beth's local manipulation case but make painstakingly clear the many different compatibilist conditions she has satisfied:

Beth is an incredibly kind woman who would not dream of hurting a fly and devotes a lot of time and money to various charitable causes. While she is sleeping, she is abducted by neurologists who erase her old values and traits that caused her to act nicely, and implant values that the scientists know will cause Beth to act cruelly. Beth is unaware of these events and has not given prior consent for them to occur. When Beth wakes up, she is surprised that she does not feel like volunteering as she normally does, but would rather murder her neighbour.

As Beth reflects on this desire, she realises that she would really like to murder her neighbour and that she wants to be the kind of person who commits crimes. Her neighbour has been irritating her and she highly values living a more peaceful, self-interested life without him around. Beth's desire is not irresistible but she chooses to act on it anyway. Beth understands that murdering one's neighbour for personal gain is morally wrong, but this does not deter her. Though she can regulate her behaviour with moral reasons, Beth finds her self-interest to be more motivating in this case. Beth is capable of feeling guilt, shame, pride, anger and a whole host of other reactive attitudes and is receptive when these emotions are directed at her. She feels some guilt knowing that the action is wrong, but her neighbour has made her angry and she feels glee at the thought of her murdering him. Beth carefully deliberates about this action, weighing up the pros and cons of the action. She does not believe that she will get caught and reasons that she stands to gain a lot from her neighbour's death. Beth is reasons-responsive; if she believed that there was a good chance she could go to jail, or if she believed the benefits of killing her neighbour were less than they currently are, then she would choose not to murder him. After this deliberation, Beth sets out and successfully murders her neighbour.

When Beth's deliberative process is described and there is emphasis placed on the number of compatibilist accounts of control that she satisfies, it is intuitive that she is responsible for

¹¹⁵ McKenna, "The Relationship between Autonomous and Morally Responsible Agency," 217.

the murder of her neighbour. She no longer seems like the tool of distant, malevolent neurologists but appears to be acting autonomously. When it is made clear that she has reasoned carefully (more carefully than we do for most of our actions) and fully endorses the action she is going to perform, we understand that Beth has the same mental capacities as you or I and should be held responsible for exercising them just as we are with our own.

Pereboom has disputed that this kind of reply is a challenge for his global manipulation argument against compatibilism. He has maintained that if there is equal weight given to the hidden causes of Beth's behaviour, as well as her agential properties, then we will have the intuition that Beth is not responsible.¹¹⁶ The hidden causes Pereboom wishes us to focus on are such things like subconscious influences on our behaviour, as well as deterministic chains of events that extend into the past, before Beth was born. Whether this tactic is successful as a defence for global manipulation arguments has been debated by McKenna.¹¹⁷ But what concerns us is if it can be used to defend historical compatibilism. I contend that it cannot.

Pereboom's reply cannot be adapted by historical compatibilists because they do not believe that the reason local manipulation negates freedom is because it is a particularly salient example of a hidden cause. As they are compatibilists, they have no qualms with someone's behaviour being the product of a long line of deterministically caused events that began long before the agent was born. Pereboom's original argument and response are attempts at eliciting the intuition that these hidden causes undermine moral responsibility.

¹¹⁶ Derk Pereboom, "Defending Hard Incompatibilism," *Midwest Studies in Philosophy* 29, no. 1 (2005); "A Hard-Line Reply to the Multiple-Case Manipulation Argument," *Philosophy and Phenomenological Research* 77, no. 1 (2008).

¹¹⁷ Michael McKenna, "A Hard-Line Reply to Pereboom's Four-Case Manipulation Argument," "Resisting the Manipulation Argument: A Hard-Liner Takes It on the Chin," *Philosophy and Phenomenological Research* 89, no. 2 (2014). See also, Daniel Haas, "In Defense of Hard-Line Replies to the Multiple-Case Manipulation Argument," *Philosophical Studies* 163, no. 4 (2013).

But historical compatibilists, in virtue of being compatibilists, have already conceded that Pereboom is incorrect on this point. As argued earlier, it is mysterious why historical compatibilists believe that manipulation is a problem for free will and moral responsibility. But it is clear that it is not due to manipulation being an example of a hidden cause.

There is a second aspect of Pereboom's reply that historical compatibilists might try to adopt. It is his belief that equal attention should be given to the manipulation and the agential properties and this will lead to a person having the intuition that Beth is not responsible. I am uncertain how a historical compatibilist would like the case to be described because when Pereboom made this point, I believe what he had in mind is a detailed description of the hidden subconscious or external events that led to Beth performing the act, but as I have argued, historical compatibilists would not want to fixate on these hidden causes. As it stands, I do not believe that historical compatibilists can protest my description of the case as unfair. The first part of the case describes the manipulation similarly to the way it has been described by historical compatibilists. The second part focuses on Beth's properties, but it does frequently mention her desire to murder, which has just been described as being implanted by neuroscientists. Given this, I do believe that historical and non-historical compatibilists should be happy with how I have described the case. Furthermore, when I read the case, my intuition is that Beth is free and responsible for the murder.

When we emphasise that manipulated agents are still very similar to ordinary agents and can deliberate just as well or even better than manipulated agents, we get the intuition that they are free and responsible for the actions they perform. This intuition is very important for defending non-historical compatibilism because it clearly shows that, when described properly, it is not intuitive that manipulation nullifies responsibility. It also supports my

previously stated point that praising and blaming have consequentialist aspects to them, because it becomes intuitive to blame Beth for what she does when we realise that Beth is able to appreciate reasons and can reflect and alter her behaviour. Moreover, Pereboom's response, which is the only extant reply to this kind of response I am aware of in the literature, cannot be used successfully by historical compatibilists. For these reasons, I believe focusing on agential properties strongly supports non-historical compatibilism.

4.7. Responsibility-preserving Manipulation

It is often assumed by all parties to this debate that manipulation undermines moral responsibility and that it is a strike against non-historical compatibilism that it denies this intuition. In Section 4.6, I presented a thought experiment where it was intuitive that Beth was praiseworthy for what she did, despite being locally manipulated. In this section, I will discuss Neal Judisch's ingenious manipulation case, which is yet another case where it is intuitive that the agent described is morally responsible though they are manipulated. I will present his thought experiment and develop it further by anticipating and criticising potential objections. Moreover, the historical compatibilist is unable to accommodate this intuition, thus giving the non-historical compatibilist an edge on this case.¹¹⁸

Imagine a person called Malory. She had an ordinary upbringing, received a moral education, and is a normal human being much like you and me. She has never been manipulated, so she does not act from compelled pro-attitudes and she has taken responsibility for the mechanisms that she acts from, thus satisfying both Mele and Fischer & Ravizza's history requirements. One night, while she is sleeping, Malory suffers a brain

¹¹⁸ The following comes from Neal Judisch, "Responsibility, Manipulation and Ownership," *Philosophical Explorations* 8, no. 2 (2005).

lesion. The lesion corrupts the signals transmitted by her neurons such that the pro-attitudes she now holds are completely different to the ones she held before, and if she were to wake up now, Malory would discover that she is no longer reasons-responsive, cannot act in line with what her deep-self values and so on.

Fortunately for Malory, Dr. Krieger is somehow made aware that one of his patients has a brain lesion. The lesion is irreparable, but he is able to insert an electronic device in Malory's brain that is able to receive the corrupted neural signals, 'translate' them, and communicate them to other regions of her brain, thus performing all the work that was normally reserved for the part of her brain that has a lesion. This operation restores Malory's capacities to what they were prior to the lesion. Krieger leaves, dies, and is never seen or heard of again. Furthermore, Krieger's device is never discovered. When Malory awakes, she is unaware that she has recently both developed a brain lesion and undergone surgery. Her life seems exactly as it was before, and she gets up and greets her neighbour like she usually does.

Intuitively, Malory is responsible for greeting her neighbour. It is what she would have done had she not suffered a lesion. Moreover, if Krieger had not inserted the machine that sends neural messages to various parts of her brain, then the acts she would have performed would have expressed very different values to the ones she previously held. But thanks to Krieger manipulating Malory's brain, we can say that her responsibility is restored or preserved.

This intuition should trouble historical compatibilists. Malory's brain has been tinkered with, in order for it to function a particular way, similar to what happened to Beth. The only difference between the two appears to be that Malory's post-lesion character and abilities are identical to her pre-lesion character and abilities, whereas Beth is quite a different person after her manipulation. But as argued in Chapter III, Section 3.7.2, if manipulation

renders agents non-responsible, then facts about differences in character do not adequately explain why. As historical compatibilists' historical requirements are motivated by cases like Beth's manipulation, they also run the risk of denying responsibility in manipulation cases like Malory's where it is intuitive that she is responsible for what she does. Non-historical compatibilists do not face the problem of having to bite the bullet in responsibility-preserving manipulation cases, as they believe considerations of manipulation are immaterial to considerations of moral responsibility, and can grant that Malory is responsible.

According to Fischer & Ravizza, Malory is not responsible for what she does. To be morally responsible on their view requires taking responsibility for the mechanism that you act from. Fischer & Ravizza do not offer a complete account of mechanisms, though when discussing a manipulation case, they do state that "in taking responsibility for ordinary practical reasoning, Judith [the victim] does *not* thereby take responsibility for a *different kind of mechanism*—one that involves direct stimulation of the brain."¹¹⁹ Given that Fischer & Ravizza do not think you take responsibility for functionally equivalent mechanisms if their internal features differ, then it can be seen that Malory has not taken responsibility for Krieger's device. Consequently, Fischer & Ravizza's positive historical compatibilism cannot accommodate the intuition in this case.

Mele holds that if an agent is morally responsible, then she does not have a history that contains her being compelled to have pro-attitudes. If Mele is to claim that Malory is responsible, then we need to determine whether she has been compelled according to Mele's theory. I will assume for the sake of argument that Malory satisfies Mele's other *NFM* conditions of compulsion (in Chapter 1, Section 1.6.1). I believe the condition that is

¹¹⁹ Fischer and Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, 234. Emphasis in original

contentious whether she satisfies it or not is the third condition. Recall that the third condition holds:

- (3) Those values were very recently produced in a way that bypassed her capacities for control over her mental life by value engineering to which she did not consent and are seriously at odds with autonomously acquired values of hers that were erased in the process.¹²⁰

Malory's current pro-attitudes are possessed because of value engineering, which did bypass her compatibilist agential structure without her consent. So we cannot look to any of her present pro-attitudes to defend the compatibility of Mele's view with the intuition that Malory is responsible. But, the condition also makes reference to non-bypassed possessed pro-attitudes that Malory held earlier. If this can be interpreted so that Malory's pro-attitudes, prior to the brain lesion, support her identifying with her current pro-attitudes and make her morally responsible for acting on them, then Mele's account will be able to dodge a potential counterexample.¹²¹

However, I do not think (3) should be understood in this way, as it becomes prone to counterexamples. On the current reading of (3), it would mean that Malory having a set of pro-attitudes at t_1 , is sufficient for her to have the right kind of history for being morally responsible when she has a similar set of pro-attitudes at t_3 , despite the fact that at t_2 , Malory possessed very different pro-attitudes. So Malory having a sunny disposition yesterday would mean that she still satisfies Mele's historical condition today when she

¹²⁰ Mele, *Free Will and Luck*, 170; "Manipulation, Compatibilism, and Moral Responsibility," 268-69; "Moral Responsibility and History Revisited," 470.

¹²¹ I am not claiming that this is what Mele himself believes, but that his account can be interpreted or amended so that it reads this way.

cheerily greets her neighbour, though she was an unpleasant person throughout most of the night.

The problem with this view is that there is a gap between the two sets of pro-attitudes, and that they are only connected by the fact that the pro-attitudes of the earlier time resemble the pro-attitudes at a later time. This problem is highlighted if we imagine that the brain lesion causes both a change in Malory's character and induces her to sleep until technology of the distant future is able to bypass it. When Krieger's descendant wakes her in a few thousand years, it would be odd to suggest that Malory is responsible for what she does because she bears resemblance to a person who fell asleep so many years ago. It would be stranger still if Krieger was unable to restore Malory's character, and so we denied that Malory is responsible because she did *not* bear resemblance to a person who fell asleep thousands of years ago.

This reading of (3) can again be seen as counterintuitive if we imagine a person who, unlike Malory, does not sleep while holding different pro-attitudes. Imagine that River was part of a cult that prized violence above all else, and that she once strongly held these values. However, once she began to mingle with people from outside of the cult, she slowly moved away from these values and eventually became a pacifist, though she is open to returning to the cult. The cult leaders, not wishing to lose a member, kidnap River one evening and insert a device into her brain, similar to Krieger's, that causes her to value violence once more. This situation is parallel to Malory's in that it has someone who once held the values, loses them, and through manipulation regains the values. River's case is different to Malory's in that it is not intuitive that River is morally responsible when she acts from her returned values. Defenders of this interpretation of Mele's theory will have to accept that individuals in positions similar to River's are morally responsible, if they wish to

claim that Malory is also morally responsible. Due to these counterexamples, I believe this reading of (3) is untenable, and that historical compatibilists must bite the bullet and say Malory is not morally responsible for what she does.

Historical compatibilists may want to argue that certain kinds of manipulation do not violate their historical requirements. That is, there is beneficial manipulation (like Malory's case) and harmful manipulation (like Beth's case), and only the latter interferes with an agent's history, whilst the former does not. But it is not clear why it is the consequences of the manipulation that matter. As argued in Chapter 3, Section 3.7.2, it seems that the problem with manipulation is that it bypasses the agent's rational faculties, not that the post-manipulation person is a bit different to the pre-manipulation person. Moreover, I do not see how a historical requirement could manage to discriminate between these cases while also not being *ad hoc*. Finally, if a historical compatibilist is able to formulate a historical condition that makes this distinction, I believe it will face counterexamples similar to the ones I presented against alternative readings of Mele's view. For these reasons, I believe the historical compatibilist should admit that they cannot capture this intuition, and that non-historical compatibilism has an advantage over it in these cases.

4.8. Mele's 'First Action' Clause

When stating the conditions for when an agent is not free, Mele includes an interesting caveat which has not been discussed in the literature but I believe merits attention. Mele's fifth criterion is:

NFM. An agent who performs an overt action *A* does not freely *A* and is not morally responsible for *A*-ing, if

(1) *A* is the first overt action she performs on the basis of his new values.¹²²

I wish to focus on the fact that Mele only specifies the first overt action and not any other subsequent actions the agent may perform. I find this focus puzzling, and any motivations for that clause's specification I believe vindicate non-historical compatibilism.

If Mele believed that whenever you acted on the basis of manipulated values, your act was never free, then I believe the condition would have reflected that fact by excluding the word 'first.'¹²³ But because he has not done that, I believe that Mele wishes to leave open the possibility that a later act, say *Z*, that is also performed on the basis of manipulated values could be a free act for which the agent can deserve credit or censure. But how could it be the case that an agent, acting on manipulated values *V* at t_1 , performs *A* unfreely, but the same agent, at a later time t_2 , performs *Z* because of the same value *V*, but does so freely? It may be thought that Mele thinks that *V* could at a later date come to be supported by another pre-existing value held by the agent, thus making *Z* a free act. However, Mele rules out this possibility with his third NFM condition and stipulates that *V* is in contrast with the recently erased values.

The most likely answer to the question I believe, is the intuition that Mele may be trying to appeal to with his fifth clause. The intuition is that after a long time of performing actions

¹²² *Free Will and Luck* (New York: Oxford University Press, 2006), 170; "Manipulation, Compatibilism, and Moral Responsibility," *The Journal of Ethics* 12, no. 3-4 (2008): 268-69; "Moral Responsibility and History Revisited," *Ethical Theory and Moral Practice* 12, no. 5 (2009): 470. Mele's conditions of compulsion were originally listed in Chapter 1, Section 1.6.1.

¹²³ Of course, Mele may believe that all actions, which satisfy his first four criteria for unfree acts, are unfree but simply chose to only single out the first act. Perhaps to more easily elicit the intuition he was trying to motivate? While this could be the case, I believe it is more likely that he would have stated that view if he had held it, rather than withhold it for some argumentative point.

on the basis of a once engineered character, the agent may come to 'authentically own' the character. Months or years after the manipulation, the agent may have met a new partner and friends, a new career, and every facet of his life may accord with the new values instilled in him. We may then be inclined to think that the values expressed by the agent's actions do not reflect just some neuroscientists' meddling, but who they deep down now really are. This view is plausible when we consider the fact that the agent will be able to reflect on the values he holds and revise his values based on various factors such as his friends and families' reactions, his understanding of morality, and empirical facts about the world around him. Given that he may adjust or amend, reflect and revise, and attenuate various aspects of his new values between t_1 and t_2 , it is not improbable that one could find it intuitive that an agent's *A-ing* is unfree but not his *Z-ing*.

However, I will argue that our judgements in this area must be uniform. If you believe this agent's *A-ing* is free or unfree, then you must evaluate their act of *Z* similarly. I do not believe it is possible to consistently hold that act *A* is unfree but *Z* is free; historical compatibilism cannot be reconciled with this intuition. This is because historical compatibilists believe, at the very least, that the first act you perform on the basis of values that bypassed your capacities is an action that is not free. The explanation for why acting from these values makes your act unfree is due to the way they were acquired. This is why action *A* is considered unfree. But if after *A*, the agent was to immediately perform action *B* from the same values, it would still be the case that the values were acquired through manipulation, so *B* is not free. And the same applies to actions *C*, *D*, and so on until *Z*, an act that could be many years after the manipulation.

Therefore, it does not seem that you can maintain that *Z* is a free act, while simultaneously insisting that *A* is unfree. Any reason to claim *A* is unfree is a reason to claim

that Z is unfree. Both are due to manipulation, which cannot be both responsibility-nullifying and responsibility-affirming. Moreover, the fact that an agent can attenuate, adjust, and reflect on their values V and alter them to V^* does not assist the historical compatibilist's argument. Reflection must be taken up from one perspective or another. As we are assuming that no non-bypassed values are to influence a change in V , then the perspective the agent must adopt will be from one of his bypassed values. And it does not seem helpful for responsibility if, on the basis of manipulated values, you alter V to V^* . Instead, it now seems that you have a value that was indirectly produced by manipulation, not directly produced, but that is not a relevant difference between the two.

Either both A and Z are unfree, or both are free. You cannot attempt what Mele endeavoured to do and leave open the possibility that one might be a free act and the other unfree. Non-historical compatibilists will argue that the intuition supporting the view that Z is free is stronger and more robust than the intuition supporting the belief that A is unfree and so we should say both are free. Historical compatibilists will state the contrary and insist that both acts are unfree. The non-historical and historical compatibilist will resolve the conflict between these two intuitions differently.

This appears to be an area of the debate between the two kinds of compatibilists that ends in a stalemate. However, this discussion was still fruitful. I have illustrated that historical compatibilists like Mele are unable to have their cake and eat it too; they must choose one of the intuitions to go after as they cannot have both. Moreover, I have highlighted the intuition that an agent can, after a long period of time, come to 'authentically own' their once manipulated values, an intuition that heretofore had not been used to defend non-historical compatibilism, thus giving it some support it had not

before received. It is another intuition we must weigh up when we are considering overall which theory we find to be most plausible.

4.9. Manipulation As Not Relevantly Different to Being Raised Naturally

Manipulated individuals appear to be unfree because they possess pro-attitudes that circumvented their control. But this intuition can be challenged if it is demonstrated that we already believe that agents are responsible for what they do, despite acting from values they played no part in creating. I will defend non-historical compatibilism by showing that local manipulation, which historical compatibilists abhor, is not relevantly different to the ordinary development of children. I will also respond to criticisms that Mele has made of Frankfurt when he took up the same position. The intuition that adults who have a normal upbringing are responsible for what they do is very strong and endorsed by both non-historical and historical compatibilists. If it can be shown that this intuition conflicts with the historical compatibilist's intuition evoked by local manipulation cases, then I believe most individuals will discard the latter intuition. This would remove any motivation for accepting a historical thesis.

When we are young, we have little to no control over the innate mental capacities we are born with and which environment we find ourselves in, and both of these factors interact to create the values and desires we possess.¹²⁴ Moreover, when we act because of these pro-attitudes, we are often thought to be praiseworthy or blameworthy. I do not know how I acquired the desire to study philosophy, but I am still regarded as responsible for decisions relating to my education. And I would still be responsible if it was discovered that my pro-

¹²⁴ Zimmerman, "That Was Then, This Is Now: Personal History Vs. Psychological Structure in Compatibilist Theories of Autonomous Agency."

attitude originated because of the particular books my parents chose to read to me before I went to bed.

Manipulated agents are akin to ordinary humans in that both have a set of pro-attitudes that they did not have a hand in creating. The major difference between the two is that manipulation is much faster than childhood development in arriving at a person with a complete psychic structure. Supposing for the moment that this is not a relevant difference, I believe we should regard the two kinds of cases as equivalent with respect to moral responsibility. And since the intuition that individuals sometimes act freely and are responsible is much stronger than the intuition that manipulated agents are not free or responsible, I believe that we should conclude that both manipulation and ordinary childhood do not negate our freedom.

A second reason to accept the similarity between manipulation and the regular processes that lead to acquiring desires is that it conforms to the intuition that Mele was trying to appeal to with his first action clause, which I discussed in Section 4.8. That intuition was that once a period of time has elapsed after an agent has been manipulated, it is intuitive to think that they now 'authentically own' their deep-selves and are responsible for what they do. That intuition supports this argument as it demonstrates that we currently view manipulation as similar to childhood. Just as we know we cannot control our genes and environment shaping who we are, yet are still responsible for the actions that express the resultant character, we also believe that we cannot control our manipulation, but believe that we can eventually be morally responsible afterwards.

Considerations like these, I believe, lead Frankfurt to state:

A manipulator may succeed, through his interventions, in providing a person not merely with particular feelings and thoughts but with a new character. That person is then morally responsible for the choices and the conduct to which having this character leads. We are inevitably fashioned and sustained, after all, by circumstances over which we have no control. The causes to which we are subject may also change us radically, without thereby bringing it about that we are not morally responsible agents.¹²⁵

I agree with Frankfurt on these points. We should view manipulation as comparable to our genes or environment; all are things we have no control over that shape who we are, but they also do not inhibit our free will or responsibility.

Mele has objected to this particular quote of Frankfurt's, which is supposed to establish that there is no relevant difference between someone who developed normally like Chuck and a manipulated person like Beth. Mele notes that we are affected by circumstances that are beyond our control, but that this does not mean we have no control at all.¹²⁶ Chuck did not suddenly become a nasty person, or have his character forced upon him. Instead, he set out with the explicit intention to become a horrible person. Beth, on the other hand, played no part in fashioning the Chuckian values she now has. As we can influence the processes that cause us to have particular pro-attitudes, Mele rejects Frankfurt's assertion.

I believe Mele has refuted the particular wording Frankfurt has used, but not the spirit of his quote. I agree that Chuck at t_2 is responsible and was in control when he decided to become the evil Chuck at t_3 . But what about the very young Chuck at t_1 ? He was impressionable and surrounded by violence which left in him a desire to replicate violence at a later date. And the non-autonomous acquisition of this desire at t_1 is what caused Chuck to exercise control at t_2 and become a psychopath at t_3 . Mele's response to Frankfurt shows that we can use our mental capacities to control who we are and build a certain character for ourselves. But this does not mean that we can create the materials of this construction

¹²⁵ Frankfurt, "Reply to John Martin Fischer," 28.

¹²⁶ Mele, *Free Will and Luck*, 172; "Manipulation, Compatibilism, and Moral Responsibility," 270; "Manipulation, Moral Responsibility, and Bullet Biting," 172-73.

ex nihilo. They often come from our childhood, a period where we did not have any mental control. Chuck at t_2 is responsible for what he does, despite his character having ultimately emerged from a period in his life that he could not control.

Mele may have the resources to oppose arguments of this kind for another reason. Mele has asserted that children do have mental capacities that can be bypassed. He argues that children as young as five believe and desire on the basis of evidence. For example, a child may desire to watch an animated film after having previously watched one and evaluated the experience as enjoyable.¹²⁷ If this is the case, then there is a crucial difference between having a normal history and being manipulated and the argument of this section collapses. If Mele is right, then the pro-attitudes we possess as adults were autonomously acquired through applying the mental capacities we had as children, which makes the pro-attitudes importantly different to manipulated pro-attitudes, which circumvent our cognitive faculties.

However, even if children as young as five desire and believe on the basis of evidence, this does not prevent there from being adults who we believe are morally responsible for pro-attitudes they did not autonomously develop. We know from examining babies that they do not possess all their deliberative capacities. This must mean that these capacities come into existence sometime after birth but before adulthood; we can assume, along with Mele, that this happens at age five. That means that any pro-attitudes instilled before the age of five are equivalent to our genes and the environment in that we do not have any say over any of these items. Recall my example of desiring to do philosophy because my parents read to me at a young age. If we assume that this desire was manifested in me at an age young enough, say at four years old, then we would have an example of a pro-attitude that I

¹²⁷ *Autonomous Agents: From Self-Control to Autonomy*, 167-68.

did not autonomously acquire, yet I am responsible for actions that express it. Therefore, Mele's assumptions about child psychology do not refute this argument.¹²⁸

A second, more vivid example of an agent who is morally responsible for what she does though she non-autonomously acquired her pro-attitudes at a young age comes from McKenna. McKenna describes the true story of Ann, who as a young child, saw her mother suffer from cancer and undergo painful chemotherapy. The experience made Ann see life as a precious gift and to live every moment, and it is these values that Ann acts from as an adult. Though she acquired the values she acts from before she had any cognitive abilities, Adult Ann seems to act freely and is morally responsible for what she does.¹²⁹ The values do not inhibit her responsibility but are the source of it. Ann's case gives us a strong intuitive reason to believe that someone can have a practically unsheddable pro-attitude that bypassed the agent's mental capacities because they did not have any mental capacities at that age, yet still be morally responsible for what they do. And this means that we naturally think of childhood as relevantly similar to manipulation. And since developing naturally is not a problem for moral responsibility, local manipulation should not be a problem for it either.

¹²⁸ For similar points, see Stefaan E. Cuypers, "The Trouble with Externalist Compatibilist Autonomy," *Philosophical Studies* 129, no. 2 (2006): 183-85; Haji and Cuypers, "Magical Agents, Global Induction, and the Internalism/Externalism Debate," 352.

¹²⁹ McKenna, "A Hard-Line Reply to Pereboom's Four-Case Manipulation Argument," 156; Derk Pereboom, "A Hard-Line Reply to the Multiple-Case Manipulation Argument," 167; Michael McKenna, "Resisting the Manipulation Argument: A Hard-Liner Takes It on the Chin," *Philosophy and Phenomenological Research* 89, no. 2 (2014): 478.

4.10. Instant Agents¹³⁰

McKenna has thought of an ingenious response to historical compatibilism. His argument centres on an 'instant agent,' an agent that is identical to another agent in terms of snapshot properties, but has only just popped into existence due to the work of a god or quantum fluctuation.¹³¹ McKenna describes Suzie Instant, an instant agent placed in a deterministic universe one day by god, who has a wide range of (false) beliefs about her past, who unsheddably values many things, and also exercises a great deal of self-control. Her first action is to either *A* or *B*. As *A* is in agreement with Suzie Instant's values, she decides to perform that action.

Suzie Instant, who was not compelled or acting from an irresistible desire, seems to be free and responsible for *A*-ing. McKenna compares Suzie Instant with Suzie Normal, who possesses all of Suzie Instant's current time-slice properties. The only difference is that Suzie Normal lives in a world that god instantiated long ago, and developed in the normal manner from a zygote. When Suzie Normal *As*, she does it for exactly the same reasons as Suzie Instant. Given that compatibilists are already committed to judging that Suzie Normal is free and responsible, and that there is not a relevant difference between the two Suzies, compatibilists should agree that Suzie Instant acts freely when she *As*.¹³²

Many compatibilists agree that normal and instant agents should be judged similarly.¹³³ I believe the argument is successful and it immediately poses a worry for positive historical

¹³⁰ The term 'instant agent' was coined in Zimmerman, "Born Yesterday: Personal Autonomy for Agents without a Past."

¹³¹ I am assuming that instant agents with mental content are a conceptual possibility. For disagreements with this assumption, see Donald Davidson, "Knowing One's Own Mind," *Proceedings and Addresses of the American Philosophical Association* 60, no. 3 (1987).

¹³² McKenna, "Responsibility and Globally Manipulated Agents," 179-81.

¹³³ Mele, *Autonomous Agents: From Self-Control to Autonomy*, 172-73; Haji and Cuypers, "Magical Agents, Global Induction, and the Internalism/Externalism Debate," 349; Mele, "Moral Responsibility and History

compatibilists. They hold that an agent must have a certain kind of history to be morally responsible. However, Suzie Instant has no history and we believe that she is morally responsible. Therefore, this is a counterexample to one brand of historicism, positive historical compatibilism.

In addition to repudiating positive historical compatibilism, McKenna believes the case of Suzie Instant lends some credence to the proposition that manipulated Beth is free and responsible. If this can be shown to be the case, then negative historical compatibilism will also be refuted. McKenna reasons that since Suzie Instant is responsible and has no history, then only non-historical properties can explain this fact and this supports non-historical compatibilism.¹³⁴ Many philosophers have objected to this argument.¹³⁵ I will not rehash their arguments here but add that the mere fact of noting that instant agents are responsible will not refute all historical compatibilisms. Negative historical compatibilism is consonant with the intuitions elicited from instant agent cases and manipulation. This is because it holds that agents do not have to have a history, but if they do, it must not be a history of a certain kind.

I believe that instant agents can form part of an argument that criticises negative historical compatibilism and I will give that argument in Section 4.11. Before that, I wish to address some methodological points that have been made about instant agents. Mele

Revisited," 466; Kristin Demetriadou, "The Soft-Line Solution to Pereboom's Four-Case Argument," *Australasian Journal of Philosophy* 88, no. 4 (2010): 615; Neil Levy, *Hard Luck: How Luck Undermines Free Will and Moral Responsibility* (Oxford: Oxford University Press, 2011), 202; McKenna, "Defending Nonhistorical Compatibilism: A Reply to Haji and Cuypers," 269-70; "Moral Responsibility, Manipulation Arguments, and History: Assessing the Resilience of Nonhistorical Compatibilism," 160-62; Ishtiyaque Haji, "Historicism, Non-Historicism, or a Mix?," *The Journal of Ethics* 17, no. 3 (2013): 191; Alfred R. Mele, "Moral Responsibility, Manipulation, and Minutelings," 160-61.

¹³⁴ McKenna, "Responsibility and Globally Manipulated Agents," 180-84; "Moral Responsibility, Manipulation Arguments, and History: Assessing the Resilience of Nonhistorical Compatibilism," 160-63; "Defending Nonhistorical Compatibilism: A Reply to Haji and Cuypers," 269-71.

¹³⁵ Haji and Cuypers, "Magical Agents, Global Induction, and the Internalism/Externalism Debate," 347-59; Haji, "Historicism, Non-Historicism, or a Mix?," 187-96; Mele, "Moral Responsibility and Agents' Histories," 175; "Moral Responsibility, Manipulation, and Minutelings," 160-66.

argues that we should not place much weight on intuitions derived from instant agent cases. We have never encountered humans that have been created *ex nihilo* or just popped into life. As they are sci-fi creatures who have only ever existed in our imagination, Mele believes that these intuitions will not be reliable.¹³⁶ In a similar vein, though without argument, McKenna states that we should prefer the intuitions of 'closer-to-real-life' cases over the intuitions from fictitious and bizarre cases.¹³⁷ These points, if correct, undermine all non-historical compatibilist arguments that borrow intuitions from the case of Suzie Instant.

However, I believe that objections of these kinds fail. Firstly, there is reason to think that intuitions about unrealistic cases are *more* reliable than intuitions about their closest real life counterparts. As Nelkin notes about farfetched scenarios, "one might argue that their unrealistic quality helps ensure that we are focused on the stipulated features, and that we aren't implicitly but unconsciously relying on background assumptions that we bring to ordinary life."¹³⁸ As they may not be influenced by irrelevant and subconscious factors, intuitions about instant agents may be more reliable in virtue of the fact that they are unrealistic.

Secondly, it may be appropriate to give less weight to instant agent cases because of their absurdity. But that means that the main historical compatibilist argument, the local manipulation argument, will become significantly weaker, as it too relies on a case that is contrary to our everyday experiences. The possibility of a neuroscientist coming to your home while you are asleep and rigging your brain so that your values are completely reversed is just as extraordinary and foreign to ordinary life as an adult brought into existence by a god. So, if we are to give less credence to our intuitions from fanciful

¹³⁶ "Moral Responsibility and Agents' Histories," 175.

¹³⁷ Michael McKenna, "Source Compatibilism and That Pesky Ability to Do Otherwise: Comments on Dana Nelkin's *Making Sense of Freedom and Responsibility*," *Philosophical Studies* 163, no. 1 (2013): 110.

¹³⁸ Dana Kay Nelkin, "Replies to Critics," *Philosophical Studies* 163, no. 1 (2013): 125.

examples, then our credence in manipulation cases should also drop. And as I regard manipulation cases as the main motivation behind historical compatibilism, pursuing this line of thought will have the consequence of making it much easier to argue for non-historical compatibilism. This argument succeeds only at the expense of weakening the main historical compatibilist argument. For these two reasons I do not think Mele and McKenna's claims about unrealistic cases are persuasive and if they are, then they hold for nearly all philosophical arguments, including manipulation arguments.

4.11. Instant Agents and Local Manipulation

In this section I will present my argument that demonstrates that there is no relevant difference between an instant agent and local manipulation cases, so that a compatibilist who believes the agent is responsible in one of these cases must believe the other is a case of responsibility, on pain of contradiction. In this section I will develop and present my argument. It is similar to Pereboom's Four-Case argument in that it progresses in a piecemeal fashion from case to case, arguing that each new case is relevantly similar to the one before. I will first present an instant agent, who we intuitively think is morally responsible for what she does, and I will move through a set of similar cases, ending with an example of local manipulation, where the negative historical compatibilist is forced to say that responsibility is not annulled. If I am correct about the argument, the upshot is that negative historical compatibilists will have to give up their historical condition because they hold that instant agents are morally responsible but locally manipulated agents are not.

Before I go through a detailed version of the argument, it is worth pointing out how similar Suzie Instant is to a locally manipulated person. Let's suppose not only that Suzie

Instant popped into existence as a fully grown adult and performed act *A* as it was in line with her values, but that the deity who fashioned her created Suzie Instant with the intention for her to perform act *A*. That is, the god wanted *A* to happen and Suzie Instant was a means to this end. I do not believe this slight adjustment to the case will or should change the original intuition that Suzie Instant is morally responsible for *A*-ing.¹³⁹

This Suzie Instant case is very similar to locally manipulating Beth. Beth is taken, has her old values erased and given a new set, so that she will perform a specific act *A*, the same act Suzie Instant performs. However, we do not intuitively think Beth is responsible. This should strike us as a surprising result. Both cases involve an agent being intentionally given a specific deep-self so that they will perform the same action, yet we regard one as responsible but the other not. Given the many similarities the cases share, it is odd that they produce divergent intuitions. Negative historical compatibilists will reply by asserting that there is a crucial difference between the two cases; Beth has a history where she developed a set of values under her own steam, whereas Suzie Instant appeared replete with her own set of values.

Instant agents are beings that are mostly (if not entirely) composed of current time-slice properties. The fact that we believe that they can deserve praise or blame teaches us something crucial about moral responsibility. Namely, it demonstrates that *when assessing an agent's moral responsibility for action A at time t, we should treat the agent as if he/she were an instant agent at time t*. If my contention that non-historical compatibilism is correct, then the italicised claim is true. Non-historical compatibilists believe snapshot properties are necessary and sufficient for moral responsibility; they are not primarily

¹³⁹ This version of the Suzie Instant case is very similar to Mele's zygote argument. Give that compatibilists are willing to say that Ernie is responsible in that scenario, I believe they will say that Suzie Instant is responsible here.

concerned with facts about history or other time-slices that may constitute the agent's personal identity. Therefore, they treat each complete set of current time-slice properties of an agent at a time as if they were an instant agent just sprung into being.

Negative historical compatibilists, on the other hand, believe that it is important that Beth had a history that has been violated by manipulators. I believe that accepting that instant agents are morally responsible means that, in order to remain consistent, you must accept that locally manipulated agents are responsible. Given that negative historical compatibilists accept that instant agents are morally responsible but not locally manipulated agents, this causes a problem for their view. The rest of this section will be devoted to arguing for the view that accepting that instant agents are responsible means that one should treat all agents as if they are instant agents, from the perspective of moral responsibility.

I will move through a series of cases, arguing that there is no relevant difference between them. The first case is as follows:

First case: Suzie Instant is a person who was created by a god at age thirty, and is indistinguishable from an adult who had actually lived the past thirty years. She has a complete psychological profile which includes a set of beliefs, values, and pseudomemories. She exercises the same amount of control over her behaviour as any other normal adult. The deity made Suzie so that she would give money to the homeless. Suzie strongly values helping those less fortunate than herself and believes donating money will help achieve this end. As it aligns with her deep-self, at t_1 Suzie Instant donates money to the homeless.

This case is just a fuller description of the Suzie Instant case I have described before. It seems intuitive that, despite her lack of history, Suzie Instant is morally responsible for her decision to donate to the homeless. This is something that both a negative historical compatibilist and a non-historical compatibilist can agree on.

The second case is slightly different to the first:

Second case: Suzie Instant is created the same as before and donates money to the homeless at t_1 , an act she believes will help the homeless, and helping the homeless is something she desires highly. Suzie Instant is then suddenly destroyed by a jealous demon. The god then creates Sandra Second, another instant agent whose mental economy is packed with beliefs and pseudomemories that are identical to Suzie's. Sandra's values are very slightly different to Suzie's as she values the plight of the women's refuge slightly higher than the homeless. The god created Sandra with the intention that she would donate money to the women's refuge. Sandra is committed to benefitting those worse off than herself and believes a financial contribution to the women's refuge will satisfy this aim. Sandra then gives money to the women's refuge at t_2 , an act that is within the control of her deep-self.

The second case seems to be just the same as the first, except now there are two independent instant agents who carry out an action. This does not seem to be a barrier to granting that both Suzie Instant and Sandra Second are morally responsible and praiseworthy for the actions they each perform.

Third case: Suzie was created the normal way and is currently 30 years old. She is identical to the Suzie's of Cases 1 and 2 in terms of her motivations and beliefs, though she has legitimate memories, not merely pseudomemories. Suzie feels very strongly about the plight of the homeless and after much consideration, she believes her best of course of action is to donate some money to the homeless, which she does at t_1 . After witnessing this kind act, an evil demon completely obliterates Suzie's brain so that it will not be done again. A god then plants a new brain inside Suzie's old body. We can dub this combination of a new brain and Suzie's old body 'Sandra.' This new brain is identical to the one possessed by Sandra Second in Case 2. This means that she has the same beliefs as Suzie, except Sandra values eradicating domestic violence slightly higher than eliminating homelessness. Moreover, Sandra has the same memories as Suzie, except these are pseudomemories because they are not causally connected to the events that actually took place. After much deliberation, Sandra decides that she would like to help the women's refuge and donates money to a charity at t_2 .

In the third case, the lives of Sandra and Suzie are much more closely connected than in the previous case. In this case, Suzie is no different to an ordinary adult woman, so we will regard her as responsible for her donation. Although it is trickier to say, I believe that it is intuitive that the new brain, which is a value twin of Sandra, is both a different person to Suzie and morally responsible for the donation at t_2 .

Because Suzie's brain is completely destroyed and replaced with another brain, I believe we will view the being I called 'Sandra' as a new person. Some philosophers hold that some

kind of physical continuity is sufficient for personal identity and might argue that it is still Suzie performing the action at t_2 . For their view to apply here though, they would have to suggest that personal identity is constituted by a body part that is not the brain, a claim that is highly counterintuitive. Though physical continuity might constitute personal identity and despite the fact that the rest of Suzie's body is intact, I believe the destruction of her brain is equivalent to the death of Suzie.¹⁴⁰

Given this, the creation of a new brain means that there is a new person, Sandra, who happens to look identical to Suzie. Since this is a new brain, created *ex nihilo* without a past, I believe we should view Sandra as another instant agent. And as we were willing to say that both the instant agents in case two were morally responsible for their donations, I believe we should say that normal adult Suzie at t_1 is responsible for her donation, and so is Sandra at t_2 .

The fourth case is a slight twist on the third case:

Fourth case: A person called Suzie, who is identical to the Suzie of Case 3, donates money to a charity for the homeless at t_1 . Outraged, an evil demon removes the sections of her brain that contain her values so that Suzie will not perform this action again. A god then attempts to repair Suzie's brain by replacing the parts that are missing. However, the new parts she receives have slightly different neurological structures, so she now values assisting the women's refuge slightly higher than assisting the homeless. She is equally as concerned and motivated to help the women's refuge as the Sandra's of Cases 1-3. Though these new values differ from the ones held at t_1 , they are not so drastically different that they cause a dissonance with Suzie's still held memories and beliefs. Suzie reflects on her values and beliefs, and then concludes that she ought to give money to the women's refuge. She does so at time t_2 .

Up until t_1 , the fourth case is identical to the third case. Therefore, we should give the same judgement as before and say that Suzie is responsible for her action at t_1 . In Case 3, I argued that when Suzie's brain was destroyed and replaced, there was a different person inside

¹⁴⁰ For a defence of the view that something physical is required for personal identity, see Bernard Williams, "The Self and the Future," *The Philosophical Review* 79, no. 2 (1970).

Suzie's body at t_2 . However, since only the value sections of Suzie's brain have been removed and replaced with something very similar to the original, I believe it is intuitive to say that in Case 4, Suzie survives the demon's lobotomy and performs the action at t_2 .

In addition, I believe that Suzie is morally responsible for the act performed at t_2 . I do not see that there is a relevant difference between Cases 3 and 4 in this respect. In Case 3, we are willing to say that replacing an entire brain does not hamper the fact that someone was responsible for the donations at t_2 . It would strike me as odd if we were to then say that replacing only parts of the brain is sufficient to inhibit Suzie from being morally responsible for her donation at t_2 in Case 4.

A second difference between the two cases is that in Case 3, Sandra at t_2 has pseudomemories, whereas in Case 4, Suzie at t_2 has actual memories because she is causally connected to the events that took place in the appropriate way. However, I do not see why this point would give us reason to question Suzie's capacity for moral responsibility here. On the contrary, it would seem that only having pseudomemories is what should inhibit our capacity for moral responsibility. But since we already found it intuitive that Sandra with her false memories is responsible, it seems that we must also find Suzie with her true memories at t_2 praiseworthy because there is no relevant difference between the two.

Some negative historical compatibilists may object at this point and argue that there is not enough psychological similarity between the people at t_1 and t_2 in Case 4 for them to be the same person. I believe that this is a plausible contention. But if it is true, I believe it leads more credence to non-historical compatibilism. We would be assuming in Case 4 that a new person has emerged at t_2 . She comes with a set of values and beliefs, and her memories are false in the sense that the person who remembers the events is not the same person who was originally involved in the events (according to the hypothesis that we are entertaining

which holds that this is a new Suzie). If we hold that this is a new person, then it is apparent to me that this person is not relevantly different to the instant agents we have been discussing; both do not have a history as they have emerged from circumstances they had no control over. And since we are willing to say instant agents are responsible for what they do, then I believe we should say that in Case 4, if there is a new person at t_2 , then she is responsible for the donation she makes.

A final point to make clear is that the hesitation we have to grant that the person at t_2 is responsible for their action rests squarely on worries about personal identity. Issues surrounding the agent's capacities, abilities, understanding of morality, or the effectiveness of praise do not emerge (which is not surprising given that we did not have similar issues with Sandra Second). And as argued above, any worries to do with personal identity can be assuaged. Either it is still Suzie at t_2 , in which case it would be odd to say she was not responsible, given our judgement of Case 3. Or it is not Suzie but a new person at t_2 , in which case it seems we should treat them analogously to an instant agent. Regardless of which position you take, you have good reason to believe the person at t_2 deserves credit for donating to the women's refuge.

Fifth case: A person called Suzie, who is identical to the Suzie of Case 3, donates money to a charity for the homeless at t_1 . A neuroscientist hired by the women's refuge then kidnaps her and programmes her so that she cares more for the hardships of women than the homeless. She cares for women as strongly as the Sandra's of cases 1-3. Once released by the neuroscientists, Suzie sees someone collecting for women's refuge. She reflects on her values and beliefs, and because she very much desires to, she donates money to the women's refuge at t_2 .

Case 5 is just a reiteration of the local manipulation examples that I have been discussing. The only substantial difference between the fourth and fifth case is that the contents of Suzie's mind have been altered, as opposed to having parts of her neurobiology physically

removed and replaced. But since the effect on her cognitive capabilities and deep-self is the same as what was done in Case 4, I do not think that the method for how this was achieved will impact our intuitions.

This result is troubling for negative historical compatibilists. Because we accepted that Suzie is responsible in Case 4 for her actions at both t_1 and t_2 , and as we do not believe there is a relevant difference between Suzie in Case 4 and Suzie in Case 5, we must say that Suzie in Case 5 is morally responsible. Which means that we cannot view local manipulation as disbaring agents from moral responsibility.

Rather than accepting this conclusion, it is possible that a negative historical compatibilist could deny that instant agents are morally responsible. This would halt the generalisation from Case 1 to Case 5 by not conceding on Case 1. Though this reply is open to them, I believe negative historical compatibilists should not accept it because then they will face the same problem as positive historical compatibilists faced in Section 4.10. That is, it is intuitive that instant agents are morally responsible for what they do and it is a cost for any position to deny that claim. Moreover, wanting to include agents with no history as responsible seems to be one of the few motivations behind accepting a negative version of the historical condition. It is unclear why one would believe negative historical compatibilism if it took the same stance on instant agents as positive historical compatibilism. For these reasons, I do not think negative historical compatibilists should deny that instant agents are morally responsible.

The consequence of this argument is that the negative historical compatibilist position is false. That is, by accepting that instant agents can be responsible for what they do, negative historical compatibilists must also accept that local manipulation does not undermine moral responsibility. And since manipulation arguments are the main motivation behind historical

conditions, there is no reason to put forward a historical condition which means we should side with non-historical compatibilism.

Chapter V – Conclusion

5.1. Summary

I have presented strong reasons to doubt the arguments that purported to support historical compatibilism. In Chapter 2, The Argument from Tracing was discussed. This argument relied on the notion of tracing being a historical phenomenon to demonstrate that our theory of moral responsibility was already historical. However, I responded that tracing was in fact a forward-looking concept and not at all similar to the historical conditions the historical compatibilist was trying to defend. In Chapters 3 and 4, I explained and evaluated the manipulation argument, which held that non-historical compatibilists are committed to the belief that manipulated agents act freely and are responsible for what they do, and that this belief is implausible. I criticised the argument by highlighting that, when examined, it was not such a counterintuitive claim to make. Furthermore, I argued that there are intuitions that could only be adequately explained under non-historical compatibilism, and that for historical compatibilism to remain a viable option, it had to make a number of *ad hoc* distinctions about processes that, on the face of it, do not seem relevantly different to the local manipulation they claim to abhor.

The upshot of all this is that I believe we now have a set of very powerful reasons for rejecting historical compatibilism and accepting non-historical compatibilism.

5.2. Moderate Historical Compatibilism?

In light of my arguments, a compatibilist may agree that historical compatibilism is incorrect, but instead of endorsing non-historical compatibilism, they may move towards a position I call *moderate historical compatibilism*. Moderate historical compatibilism is an attempted synthesis of the two positions I have discussed. Moderate historical compatibilists agree with non-historical compatibilists that agents who exercise control over their actions are morally responsible for them, but they also concur with historical compatibilists by believing that historical facts can influence moral responsibility. The aim of the position is to recognise the merits of the arguments in favour of non-historical compatibilism, whilst also retaining the intuitions that support historical compatibilism without taking on the position's flaws.

Moderate historical compatibilism holds that moral responsibility comes in degrees. If you have the right history and exercise control over an action, then you have full moral responsibility for that action. But if you have the wrong history and exercise control over an action, you are still responsible for that action, but you are less responsible than if you had had the right kind of history. Historical compatibilists have viewed moral responsibility as a binary concept; either you have the right history and you are responsible, or you do not have the right history and you are not responsible. Moderate historical compatibilism conceptualises moral responsibility as scalar; you are responsible if you exercise control, but how praiseworthy or blameworthy you are depends on historical facts.

To illustrate, consider how a moderate historical compatibilist would respond to the Chuck and Beth case (first presented in Chapter III, Section 3.5). Since Chuck has the right history and intentionally became a reprehensible person, he is fully blameworthy for his

murder and deserves 10 years in prison.¹⁴¹ Locally manipulated Beth, on the other hand, does not have the right history but still exercised control when she murdered, so she is blameworthy, but only deserves 5 years (or some number less than 10). Moderate historical compatibilists hold Beth responsible, because they recognise, along with non-historical compatibilists, that she was in control when she acted and her deep-self is as corrupted as Chuck's. But they also believe, in the vein of the non-historical compatibilist, that it is no surprise she has ended up that way, given her local manipulation. She is a victim of her circumstances and this historical fact mitigates her responsibility and makes her less blameworthy than Chuck.

Though I find moderate historical compatibilism to be a more palatable position, I believe it faces the same problems as historical compatibilism. The fault with moderate historical compatibilism is that it enshrines the same view of history as historical compatibilism; by accepting that history can mitigate moral responsibility, moderate historical compatibilism must admit that there is something problematic about certain kinds of histories. But to successfully hold this view, it must counter some of the arguments I gave in Chapter IV.

Moderate historical compatibilists must find a non-*ad hoc* way to distinguish between bad histories (such as local manipulation) and good histories (such as responsibility-preserving manipulation and being raised naturally) which I do not believe is possible.¹⁴² They must also explain why history mitigates responsibility, even though when we concentrate on the agential properties of a manipulated person, we find it intuitive that they are as responsible as an ordinary person.¹⁴³ They will need to combat my argument

¹⁴¹ The number of years in prison are only meant to demonstrate how the moderate historical compatibilist conceives of moral responsibility and are not supposed to be indicative of what punishment a murderer should receive.

¹⁴² This is the argument of Chapter IV, Sections 4.7 and 4.9.

¹⁴³ The argument of Chapter IV, Section 4.6.

that we have good consequentialist reasons for believing manipulated agents are just as responsible as normal agents.¹⁴⁴ Finally, we do not think that an instant agent's lack of history mitigates their responsibility. This fact can be used in an argument, similar to the one of Chapter IV, Section 4.11, to argue that there is no relevant difference between an instant agent and a manipulated agent, so that their history does not mitigate responsibility either.¹⁴⁵

For these reasons, I do not believe we should accept moderate historical compatibilism. Non-historical compatibilism is still the preferable compatibilist theory of moral responsibility. I will now explore the implications of accepting non-historical compatibilism by returning to Robert Harris.

5.3. The Case of Robert Harris

Robert Harris, first discussed in Chapter 1, Section 1.1, is the murderer who killed two 16 year old boys in cold blood. Given how callously Harris was and his jovial reaction in response to the boys' death, it seems his deep-self wholly endorsed his inhuman actions. This gives us reason to think he is responsible for the murders. However, Harris also had a dreadful childhood. His father was a violent alcoholic who molested Harris' sisters. His mother never showed Harris any intimacy and eventually became an alcoholic herself.

¹⁴⁴ The argument of Section 4.5.

¹⁴⁵ In Chapter IV, Section 4.3, I argued that historical compatibilism is *ad hoc* because it views local manipulation as a threat to moral responsibility but not global manipulation. Interestingly, moderate historical compatibilism may be able to overcome this objection by holding that both global and local manipulation are historical facts that mitigate moral responsibility. This is a potential advantage moderate historical compatibilism has over historical compatibilism. However, there are arguments that this position would just collapse into incompatibilism. See Patrick Todd, "A New Approach to Manipulation Arguments," *Philosophical Studies* 152, no. 1 (2011). For replies, see Hannah Tierney, "A Maneuver around the Modified Manipulation Argument," *Philosophical Studies* 165, no. 3 (2013); Andrew C. Khoury, "Manipulation and Mitigation," *Philosophical Studies* 168, no. 1 (2014).

When in a juvenile prison, Harris was raped multiple times and twice tried to kill himself. His history intuitively disqualifies him as counting as a responsible agent.

What should we say about Harris then, given that our intuition is that his tragic childhood should exclude him from guilt, yet non-historical compatibilists seem to believe that an agent's history should be dismissed from our considerations? It is important that non-historical compatibilists develop a robust reply to this question because Harris' case is similar to many other people in criminal justice systems around the world; how the non-historical compatibilist treats Harris will inform how we should evaluate the responsibility of other criminals. The answer to this question will better our understanding of non-historical compatibilism, though how we answer it depends on the assumptions we make. These assumptions are about what effects Harris' past had on his agency, and whether those effects prevent him from satisfying our best theory of freedom and responsibility.

If we assume that Harris does not satisfy our best theory, then a non-historical compatibilist will deny that Harris is blameworthy for the murders. This is not surprising because we are stipulating that Harris is not in control of his actions. But it is important to understand how Harris' history may have caused him to fail to fit the criteria of our best theory. Studies have found significant relationships between being physically abused as a child and being a violent person, as well as having emotional problems including depression, anxiety and psychosis.¹⁴⁶ Additionally, studies have found that physically abused children are more likely to attribute hostile intentions to others when most persons would not make

¹⁴⁶ Robin Malinosky-Rummell and David J. Hansen, "Long-Term Consequences of Childhood Physical Abuse," *Psychological Bulletin* 114, no. 1 (1993).

those attributions, can more readily access retaliatory aggressive responses, and are more likely to see aggressive behaviours as morally permissible.¹⁴⁷

These studies suggest that Harris and others who perpetuate violence after being the victim of it do not control their actions in the appropriate way. The information-processing biases as well as the emotional problems suggest Harris is not a reliable deliberator and does not fully comprehend the moral dimension to his actions. If this is true then Harris is not responsible, nor did he have free will for acting the way he did. This conclusion concurs with the opinion of non-historical compatibilists. Non-historical compatibilism is in accord with our pre-theoretical response.

Furthermore, Harris' history, in this interpretation of the case, is taken into account by the non-historical compatibilist. This is because past facts about Harris instruct us on what abilities we expect him to possess currently. Harris' history informs us that he probably does not have the right kind of control over his actions. This does not commit us to a historical thesis because these facts are only apparently historical, not genuinely historical. Harris' past is important in this situation only because it has causally influenced his behaviour; it directly bears on whether he satisfies the compatibilist's control condition. If Harris' past did not affect his ability to control his actions, then we could ignore it. The distinction between genuine and apparently historical facts allows the non-historical compatibilist to make judgements of Harris' responsibility based on past facts about him, without becoming a historical compatibilist.

Though this is speculation, it seems to be a good thing that, when we do learn of an agent's horrible history, it evokes sympathy in us and lessens our desire to blame people like Harris. The sympathy in these real world cases appears to be truth-tracking; we feel

¹⁴⁷ Kenneth A. Dodge et al., "Social Information-Processing Patterns Partially Mediate the Effect of Early Physical Abuse on Later Conduct Problems," *Journal of Abnormal Psychology* 104, no. 4 (1995).

sympathy for people like Harris because of his history, which diminishes our desire to blame and do him harm, and in actuality, his unfortunate history does indeed mean that he is not an appropriate candidate for blame. In most real world cases, non-historical compatibilism is able to capture our untutored reactions because our histories do give us evidence about our current-day abilities.

However, in the unlikely event that Harris did in fact have the necessary kind of compatibilist powers when he murdered those two boys, then a non-historical compatibilist will agree that he is morally responsible for them.¹⁴⁸ In this version of the case, Harris' history does not appear to have inhibited his mental capacities. I do not think it is controversial to hold Harris responsible for what he has done in this scenario. One need only remember all the different compatibilist powers Harris could have possessed (listed in Chapter IV, Section 4.6) at the time of the murder. If he satisfied our best compatibilist theory, then Harris would have deliberated carefully about the murder, his deep-self would have sanctioned his terrible crimes and so on. He would have knowingly done wrong and I do not believe that anyone thinks blame is an unnatural response to this type of action.

For these reasons, I believe that non-historical compatibilism gives the right result, regardless of what the actual truth is about Harris' particular mental states and capacities at the time. Non-historical compatibilism is able to capture a large number of our intuitions over a wide variety of cases. I do not believe the arguments for historical compatibilism are effective, nor do I believe an adequate explanation can be given for why an 'incorrect' history is incorrect. The flaws in historical compatibilism and moderate historical compatibilism lead me to accept the position of non-historical compatibilism. And its

¹⁴⁸ It does not follow that a non-historical compatibilist agrees with Harris' death sentence though. The ethics of punishment is a separate issue that can be compartmentalised and which I remain agnostic about.

application to the case of Robert Harris is illustrative that the theory can be utilised successfully.

Bibliography

- Arpaly, Nomy. *Unprincipled Virtue*. New York: Oxford University Press, 2003.
- Bealer, George. "The Incoherence of Empiricism." *Proceedings of the Aristotelian Society* 66 (1992): 99-138.
- Blumenfeld, David. "Freedom and Mind Control." *American Philosophical Quarterly* 25, no. 3 (1988): 215-27.
- Burgess, Anthony. *A Clockwork Orange*. New York: Heinemann, 1962.
- Clarke, Randolph. *Libertarian Accounts of Free Will*. Oxford: Oxford University Press, 2003.
- Conway, David A. "Capital Punishment and Deterrence: Some Considerations in Dialogue Form." *Philosophy and Public Affairs* 3, no. 4 (1974): 431-43.
- Cummins, Robert. "Reflections on Reflective Equilibrium." In *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*, edited by Michael Raymond DePaul and William M. Ramsey, 113-27. Lanham: Rowman & Littlefield, 1998.
- Cuypers, Stefaan E. "The Trouble with Externalist Compatibilist Autonomy." *Philosophical Studies* 129, no. 2 (2006): 171-96.
- Davidson, Donald. "Knowing One's Own Mind." *Proceedings and Addresses of the American Philosophical Association* 60, no. 3 (1987): 441-58.
- Demetriou, Kristin. "The Soft-Line Solution to Pereboom's Four-Case Argument." *Australasian Journal of Philosophy* 88, no. 4 (2010): 595-617.
- Dennett, Daniel. *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge, MA: The M.I.T. Press, 1984.
- Dodge, Kenneth A., Gregory S. Pettit, John E. Bates, and Ernest Valente. "Social Information-Processing Patterns Partially Mediate the Effect of Early Physical Abuse on Later Conduct Problems." *Journal of Abnormal Psychology* 104, no. 4 (1995): 632-43.
- Fischer, John Martin. "The Cards That Are Dealt You." *The Journal of Ethics* 10, no. 1/2 (2006): 107-29.
- . "Excerpts from John Martin Fischer's Discussion with Members of the Audience." *Journal of Ethics* 4, no. 4 (2000): 408-17.
- . *My Way: Essays on Moral Responsibility*. New York: Oxford University, 2006.
- . "Response to Kane, Pereboom, and Vargas." In *Four Views on Free Will*, edited by John Martin Fischer, Robert Kane, Derk Pereboom and Manuel Vargas, 184-90. Malden, Mass.: Blackwell Publishers, 2007.
- . "Responsibility and Manipulation." *The Journal of Ethics* 8, no. 2 (2003): 145-77.
- Fischer, John Martin, and Mark Ravizza. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press, 1998.
- Frankfurt, Harry G. "Alternate Possibilities and Moral Responsibility." *The Journal of Philosophy* 66, no. 23 (1969): 829-39.
- . "Freedom of the Will and the Concept of a Person." *The Journal of Philosophy* 68, no. 1 (1971): 5-20.
- . "Reply to John Martin Fischer." In *Contours of Agency: Essays on Themes from Harry Frankfurt*, edited by Sarah Buss and Lee Overton, 27-31. Cambridge, Mass.: MIT Press, 2002.
- . "Three Concepts of Free Action." *Proceedings of the Aristotelian Society* 49 (1975): 113-25.
- Goldman, Alvin I. "Philosophical Intuitions: Their Target, Their Source, and Their Epistemic Status." *Philosophische Studien* 74 (2007): 1-26.
- Greene, Joshua, and Jonathan Cohen. "For the Law, Neuroscience Changes Nothing and Everything." *Philosophical Transactions of the Royal Society of London: Biological Sciences* 359, no. 1451 (2004): 1775-85.
- Haag, Ernest Van Den. "On Deterrence and the Death Penalty." *The Journal of Criminal Law, Criminology and Police Science* 60, no. 2 (1969): 141-47.

- Haas, Daniel. "In Defense of Hard-Line Replies to the Multiple-Case Manipulation Argument." *Philosophical Studies* 163, no. 4 (2013): 797-811.
- Haji, Ishtiyaque. "Historicism, Non-Historicism, or a Mix?" *The Journal of Ethics* 17, no. 3 (2013): 185-204.
- . *Moral Appraisability*. New York: Oxford University Press, 1998.
- Haji, Ishtiyaque, and Stefaan E. Cuypers. "Libertarian Free Will and Cnc Manipulation." *Dialectica* 55, no. 3 (2001): 221-38.
- . "Magical Agents, Global Induction, and the Internalism/Externalism Debate." *Australasian Journal of Philosophy* 85, no. 3 (2007): 343-71.
- Harrison, Gerald K. "A Challenge for Soft Line Replies to Manipulation Cases." *Philosophia* 38, no. 3 (2010): 555-68.
- Inwagen, Peter van. *An Essay on Free Will*. Oxford: Clarendon Press, 1983.
- . "The Incompatibility of Free Will and Determinism." *Philosophical Studies* 27, no. 3 (1975): 185-99.
- Judisch, Neal. "Responsibility, Manipulation and Ownership." *Philosophical Explorations* 8, no. 2 (2005): 115-30.
- Kane, Robert. *Free Will and Values*. Albany: State University of New York Press, 1985.
- . *The Significance of Free Will*. New York: Oxford University Press, 1996.
- Kapitan, Tomis. "Autonomy and Manipulated Freedom." *Philosophical Perspectives* 34, no. 14 (2000): 81-104.
- Khoury, Andrew C. "Manipulation and Mitigation." *Philosophical Studies* 168, no. 1 (2014): 283-94.
- Knobe, Joshua, and Shaun Nichols. "An Experimental Philosophy Manifesto." In *Experimental Philosophy*, edited by Joshua Knobe and Shaun Nichols. New York: Oxford University Press, 2008.
- Levy, Neil. *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford: Oxford University Press, 2011.
- Levy, Neil, and Michael McKenna. "Recent Work on Free Will and Moral Responsibility." *Philosophy Compass* 4, no. 1 (2009): 96-133.
- Locke, Don. "Three Concepts of Free Action." *Proceedings of the Aristotelian Society* 49 (1974): 95-112.
- Malinosky-Rummell, Robin, and David J. Hansen. "Long-Term Consequences of Childhood Physical Abuse." *Psychological Bulletin* 114, no. 1 (1993): 68-79.
- Matheson, Benjamin. "Compatibilism and Personal Identity." *Philosophical Studies* 170, no. 2 (2014): 317-34.
- McKenna, Michael. "Compatibilism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta <http://plato.stanford.edu/entries/compatibilism/>; 2009.
- . "Defending Nonhistorical Compatibilism: A Reply to Haji and Cuypers." *Philosophical Issues* 22, no. 1 (2012): 264-80.
- . "A Hard-Line Reply to Pereboom's Four-Case Manipulation Argument." *Philosophy and Phenomenological Research* 77, no. 1 (2008): 142-59.
- . "Moral Responsibility, Manipulation Arguments, and History: Assessing the Resilience of Nonhistorical Compatibilism." *The Journal of Ethics* 16, no. 2 (2012): 145-74.
- . "The Relationship between Autonomous and Morally Responsible Agency." In *Personal Autonomy: New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy*, edited by James Stacey Taylor, 205-34. Cambridge: Cambridge University Press, 2005.
- . "Resisting the Manipulation Argument: A Hard-Liner Takes It on the Chin." *Philosophy and Phenomenological Research* 89, no. 2 (2014): 467-84.
- . "Responsibility and Globally Manipulated Agents." *Philosophical Topics* 32 (2004).

- . "Source Compatibilism and That Pesky Ability to Do Otherwise: Comments on Dana Nelkin's *Making Sense of Freedom and Responsibility*." *Philosophical Studies* 163, no. 1 (2013): 105-16.
- Mele, Alfred R. *Autonomous Agents: From Self-Control to Autonomy*. New York: Oxford University Press, 1995.
- . *Free Will and Luck*. New York: Oxford University Press, 2006.
- . "Manipulation, Compatibilism, and Moral Responsibility." *The Journal of Ethics* 12, no. 3-4 (2008): 263-86.
- . "Manipulation, Moral Responsibility, and Bullet Biting." *The Journal of Ethics* 17, no. 3 (2013): 167-84.
- . "Moral Responsibility and Agents' Histories." *Philosophical Studies* 142, no. 2 (2009): 161-81.
- . "Moral Responsibility and History Revisited." *Ethical Theory and Moral Practice* 12, no. 5 (2009): 463-75.
- . "Moral Responsibility, Manipulation, and Minutelings." *The Journal of Ethics* 17, no. 3 (2013): 153-66.
- Nagel, Jennifer. "Intuitions and Experiments: A Defense of the Case Method in Epistemology." *Philosophy and Phenomenological Research* 85, no. 3 (2012): 495-527.
- Nelkin, Dana Kay. *Making Sense of Freedom and Responsibility*. Oxford: Oxford University Press, 2011.
- . "Replies to Critics." *Philosophical Studies* 163, no. 1 (2013): 123-31.
- Orwell, George. 1984. New York: Secker & Warburg, 1949.
- Parfit, Derek. *Reasons and Persons*. Oxford: Oxford University Press, 1984.
- Pereboom, Derk. "Defending Hard Incompatibilism." *Midwest Studies in Philosophy* 29, no. 1 (2005): 228-47.
- . "Determinism Al Dente." *Noûs* 29, no. 1 (1995): 21-45.
- . *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press, 2014.
- . "A Hard-Line Reply to the Multiple-Case Manipulation Argument." *Philosophy and Phenomenological Research* 77, no. 1 (2008): 160-70.
- . *Living without Free Will*. Cambridge: Cambridge University Press, 2001.
- Pust, Joel. "Against Explanationist Scepticism Regarding Philosophical Intuitions." *Philosophical Studies* 106, no. 3 (2001): 227-58.
- Ravizza, Mark. "Semi-Compatibilism and the Transfer of Non-Responsibility." *Philosophical Studies* 75, no. 1/2 (1994): 61-93.
- Rawls, John. *A Theory of Justice*. Cambridge, Mass.: Harvard University Press, 1971.
- Shabo, Seth. "Uncompromising Source Incompatibilism." *Philosophy and Phenomenological Research* 80, no. 2 (2010): 349-83.
- Skinner, B.F. *Walden Two*. New York: Macmillan, 1958.
- Slote, Michael A. "Understanding Free Will." *The Journal of Philosophy* 77, no. 3 (1980): 136-51.
- Strawson, Galen. *Freedom and Belief*. Oxford: Clarendon Press, 1986.
- Strawson, P.F. "Freedom and Resentment." *Proceedings of the British Academy* 48 (1962): 1-25.
- Stump, Eleonore. "Control and Causal Determinism." In *Contours of Agency: Essays on Themes from Harry Frankfurt*, edited by Sarah Buss and Lee Overton. Cambridge, Mass.: MIT Press, 2002.
- Stump, Eleonore, and John Martin Fischer. "Transfer Principles and Moral Responsibility." *Philosophical Perspectives* 34, no. 14 (2000): 47-55.
- Swain, Stacey, Joshua Alexander, and Jonathan M. Weinberg. "The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp." *Philosophy and Phenomenological Research* 76, no. 1 (2008): 138-55.
- Taylor, Richard. *Metaphysics*. Fourth ed. Englewood Cliffs, N.J.: Prentice-Hall, 1974.
- Tierney, Hannah. "A Maneuver around the Modified Manipulation Argument." *Philosophical Studies* 165, no. 3 (2013): 753-63.

- Todd, Patrick. "A New Approach to Manipulation Arguments." *Philosophical Studies* 152, no. 1 (2011): 127-33.
- Vargas, Manuel. *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press, 2013.
- . "On the Importance of History for Responsible Agency." *Philosophical Studies* 127, no. 3 (2006): 351-82.
- Vihvelin, Kadri. "Arguments for Incompatibilism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta <http://plato.stanford.edu/archives/spr2011/entries/incompatibilism-arguments/>; 2007.
- Waller, Bruce N. "Free Will Gone out of Control." *Behaviourism* 16, no. 2 (1988): 149-57.
- Warfield, Ted A. "Determinism and Moral Responsibility Are Incompatible." *Philosophical Topics* 24, no. 2 (1996): 215-26.
- Watson, Gary. "Reason and Responsibility." *Ethics* 111, no. 2 (2001): 374-94.
- . "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme." In *Agency and Answerability: Selected Essays*, edited by Gary Watson. Oxford: Oxford University Press, 2004.
- . "Soft Libertarianism and Hard Compatibilism." *The Journal of Ethics* 3, no. 4 (1999): 351-65.
- . "Some Worries About Semi-Compatibilism: Remarks on John Fischer's the Metaphysics of Free Will." *Journal of Social Philosophy* 29, no. 2 (1998): 135-43.
- Weinberg, Jonathan M., Haun Nichols, and Stephen Stich. "Normativity and Epistemic Intuitions." *Philosophical Topics* 29, no. 1/2 (2001): 429-60.
- Williams, Bernard. "The Self and the Future." *The Philosophical Review* 79, no. 2 (1970): 161-80.
- Wolf, Susan. "Sanity and the Metaphysics of Responsibility." In *Responsibility, Character and the Emotions*, edited by Ferdinand Schoemann, 46-62. Cambridge: Cambridge University Press, 1987.
- Zimmerman, David. "Born Yesterday: Personal Autonomy for Agents without a Past." *Midwest Studies in Philosophy* 23, no. 1 (1999): 236-66.
- . "That Was Then, This Is Now: Personal History Vs. Psychological Structure in Compatibilist Theories of Autonomous Agency." *Noûs* 37, no. 4 (2003): 638-71.