Personal Identity and Manipulation Arguments

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy in the Faculty of Humanities

2014

Benjamin Matheson

School of Social Sciences
## Contents

1: Introduction
- 1.1 Background .................................................. 8
- 1.2 Preliminaries .................................................. 17
- 1.3 Methodology .................................................. 22
- 1.4 Outline ........................................................ 23

### Part One: Compatibilism and Personal Identity

2: Against Historicism
- 2.1 Introduction .................................................. 27
- 2.2 Structuralism .................................................. 29
- 2.3 Other potential motivations for historicism .............. 36
  - 2.3.1 Tracing cases ............................................. 36
  - 2.3.2 The development of moral agents .................... 39
  - 2.3.3 The circumstances of development .................. 40
- 2.4 Historicism and personal identity ....................... 41
- 2.5 Historicism and the psychological approach ............ 46
- 2.6 Historicism and the biological approach ............... 50
- 2.7 The collapse of historicism ................................ 53
- 2.8 Conclusion ................................................... 62

3: Authenticity without History
- 3.1 Introduction .................................................. 63
- 3.2 Authenticity, history, and narrativity ..................... 64
- 3.3 Three problems with narrativity .......................... 74
- 3.4 Brainwashed Beth ............................................ 78
- 3.5 Objection ....................................................... 81

4: Diachronic Ownership and Brainwashing
- 4.1 Introduction .................................................. 84
- 4.2 Historicism and diachronic ownership .................... 85
- 4.3 The locus of moral responsibility ......................... 87
- 4.4 Diachronic ownership and the biological approach ..... 89
- 4.5 Diachronic ownership and the psychological approach 91
- 4.6 Diachronic ownership and psychological connectedness 99
- 4.7 Narrative connectedness .................................... 105
- 4.8 Back to Brainwashed Beth ................................ 112
- 4.9 Conclusion ................................................... 118

### Part Two: Compatibilism and Manipulation Arguments

5: Manipulation Arguments against Compatibilism
- 5.1 Introduction .................................................. 121
- 5.2 The Four-Case Argument .................................... 124
- 5.3 Manipulation arguments to the best explanation ....... 130
- 5.4 Covert control is the best explanation ................... 139
- 5.5 Why is Plum not morally responsible? ................. 142
- 5.6 The redundancy of the inference to the best explanation 149
5.7 The control argument 153
5.8 Conclusion 158

6: Hard Compatibilism and Threatening Covert Control 160
6.1 Introduction 161
6.2 Threatening and unthreatening covert control 164
6.3 Soft compatibilism vs. sporadic continuous manipulation 177
6.4 Hard compatibilism to the rescue 177

7: Hard Compatibilism and Unthreatening Covert Control 185
7.1 Introduction 186
7.2 Hard compatibilism vs. unthreatening covert control 190
7.3 Hard vs. medium compatibilism 198
7.4 Covert control and moral standing 208
7.5 The collapse of source incompatibilism 218
7.6 Conclusion 221

Bibliography 221

Word count (main text only): 66040
Word count (including footnotes and bibliography): 78624
Abstract

The University of Manchester
Benjamin Matheson
Doctor of Philosophy
Personal Identity and Manipulation Arguments
July 2014

In this thesis, I defend compatibilism from all manipulation arguments. Manipulation arguments are supported by control cases. These cases purport to be counter-examples to all plausible current compatibilist conditions on moral responsibility. Some compatibilists – historicists – have argued that manipulation arguments can be undermined by endorsing the view that an individual’s moral responsibility for her actions is, in some sense, sensitive to her history. In Part One, I first argue that historicism is without motivation and is untenable. I then sketch a form of compatibilism – the structural-narrative view. This view differs from standard compatibilist accounts because it not only makes clear the synchronic ‘ownership’ (the free will or control condition), but also the diachronic ‘ownership’ conditions (normally taken to be personal identity) on moral responsibility. Both conditions have a narrative component, which I draw from narrative views of personal identity. These conditions insulate my structural-narrative from the manipulation arguments that motivate historicism, thereby providing compatibilists with a tenable alternative to historicism.

In Part Two, I argue that the remaining manipulation arguments do not show that compatibilism is false. I first clarify the structure of manipulation arguments. In particular I argue that compatibilists ought to focus their efforts on showing that the control cases that support manipulation arguments are not in fact counter-examples to the compatibilist conditions on moral responsibility. I then distinguish two types of control case: threatening and unthreatening. I argue that the remaining threatening control cases only seem to be counter-examples because of ambiguities in their descriptions that result in us misidentifying the locus of moral responsibility in those cases; once these ambiguities are clarified, the non-responsibility judgement elicited by those cases soon dissipates. I then present three related to arguments to support the claim that unthreatening cases are not counter-examples the compatibilist conditions on moral responsibility; hence that manipulation arguments do not show that compatibilism is false.
Declaration

I declare that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on Presentation of Theses.
Acknowledgements

I would first like to thank my parents, David Matheson and Catherine Matheson, for their support and encouragement throughout my studies. Before coming to Manchester, I completed two years of my PhD at the University of Birmingham. As such, I would also like to thank my Birmingham postgraduate friends for helpful discussions during my time there and afterwards: in particular, Khai Wager, Paul Broadbent, Ben Bessey, Naomi Thompson, Kirk Surgener, Sarah-Louise Johnson, and Andrew Wright. I would also like to thank the Manchester postgraduate community for their feedback during my time here. Special thanks go to Natalie Ashton for her support, encouragement, insight, her criticisms and (most importantly) for putting up with me. I owe extra special thanks to Helen Beebee for her continual support, encouragement and for helping to develop my philosophical (and grammatical) skills. I must also mention a formative experience during my studies: when I was selected to respond to Alfred Mele’s paper at a workshop on manipulation arguments in Budapest, Hungary in June 2012. I would like to thank the organisers and the participants of the workshop for this opportunity. My response to Mele eventually turned into my forthcoming Philosophical Studies paper, ‘Compatibilism and Personal Identity’ (though little to nothing of the original reply remains in the published version.) Parts of that paper (in a developed form) are featured throughout Part One (Chapters 2-4) of this thesis, though the bulk of that paper is found in Chapter 2 (§2.4-2.6). I’d also like to thank David Shoemaker – whom I later discovered to be a reviewer for my paper – for his comments. Finally, I’d like to thank my examiners, Paul Russell and Katharine Able, and the following people for last minute proofreading: Natalie Ashton, Joey Pollock, Lee Wittington, and Ben Sworn.
Chapter 1: Introduction

This thesis aims to defend compatibilism – the thesis that free will and moral responsibility are compatible with the truth of causal determinism – from all manipulation arguments.

1.1 Background

Most people think they have free will. But (1) what does it mean to say we have free will? And (2) why is it important that we have free will? Perhaps the most common answer to (1) is the following: an individual has free will if she could have done otherwise than she actually did at any given time. For example, suppose Walter bought an apple at some particular time (call it ‘t$_1$’). If it were true that he could have done otherwise at t$_1$, then he must been able to do something other than buy the apple, such as not buy the apple. Identifying the ability to do otherwise with free will is deeply entrenched in our thinking. Kleptomaniacs, for instance, have a compulsion to steal; if they steal due to this compulsion, their acts of stealing are usually considered unfree acts.

Thus compulsion is taken to be a factor that undermines an individual's free will. This presents a natural worry for compatibilists. According to compatibilism, free will is compatible with the truth of causal determinism – that is, the thesis that all facts about the past and the laws of nature together causally ensure only one future.\(^1\) But this means that if causal determinism is true then all actions are the product of the past and the laws of nature, so an individual cannot do otherwise than she actually does. Incompatibilists might then argue that if Walter cannot do otherwise than he actually does, then he is effectively compelled to act as he does.\(^2\)

---

\(^1\) This definition is based on McKenna's (2012c: 13, n. 8). Exactly what ‘causal determinism’ amounts to is controversial. My definition assumes a ‘governing’ or anti-Humean conception of the laws of nature. I do not wish to enter into the debate over the correct conception of the laws of nature. I do not wish to enter into the debate over the correct conception of the laws of nature in this thesis, so I will simply assume, in agreement with the majority of the literature, that a necessitarian or anti-Humean conception of the laws of nature is true. For more on the laws of nature and the free will debate, see Beebee (2000), and Beebee and Mele (2002).

\(^2\) A related worry for theists is this: if God exists, then – in his capacity as an omniscient being – he must know all facts about the future. But if God knows that Walter will buy an apple, it seems that Walter cannot do otherwise than buy an apple; thus it seems that Walter does not buy the apple freely. Theists respond in one of two ways: they either argue that Walter’s free will is not undermined by God’s omniscience, or that God’s omniscience is, in some way, limited as a result of giving individual's free will. Plantinga (1974) takes the first strategy. He argues that God has ‘middle knowledge’ – that is, he knows all facts about what individuals do or might do. Thus, individuals can still do otherwise in a way that preserves their free will. Todd (2013a) and Swinburne (1998) take the second strategy.
Then again, compatibilists – dating back to Hobbes and Hume – have long appealed to the distinction between compulsion and causation in their defences of compatibilism. They argue that just because an individual is caused to act by antecedent events does not mean she was compelled by those events; that is, there is a difference between an individual stealing a hat because of a psychological compulsion and an individual being causally determined to steal a hat. Even if it turns out that causal determinism is true, it seems we can distinguish the individual who is compelled to act from the individual who is simply caused to act.

Compatibilists, however, owe us an account of what counts as compelling cause and what does not. Classical compatibilists, such as Hobbes and Hume, were happy for a compelling cause to count as anything outside the individual. However, psychological disorders like kleptomania undermine this account: such psychological disorders show that compelling causes can have a source within an individual. This has prompted many contemporary compatibilists – most notably Frankfurt (1971) – to propose accounts of free will that bring the distinction between causation and compulsion (in effect) within the individual. But incompatibilists have typically been unhappy with these accounts; as they see it, individuals who are causally determined are compelled because they cannot do otherwise than they actually do. The Consequence Argument is the best expression of this incompatibilist worry:

If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born.

---

They argue that an upshot of (libertarian) free will is that God cannot in principle know the outcome of an exercise of free will. So God cannot know whether or not Walter will buy the apple. To be fair, Swinburne (1998) restricts this limitation of God’s knowledge to moral actions. Either way, I find that both theistic strategies are unpromising. But I won’t discuss these theistic strategies in this thesis.

See Russell (1988) for a discussion of this.

Again, see Russell (1988).

When formalised, this argument is known as the ‘indirect argument’. van Inwagen (1983) proposed three versions of the indirect argument, though each is thought to be independently problematic because it relies on controversial transfer principles – most notorious is the ‘beta’ principle. Fischer (1994), however, argues that the consequence argument still has force when construed without transfer principles.
and neither is it *up to us* what the laws of nature are. Therefore, the consequences of these things are not *up to us*. (van Inwagen 1983: 16; my emphasis)

That things are ‘up to us’, then, is taken to be required for free will, and for things to be ‘up to us’ it is necessary that we have the ability to do otherwise. If it is ‘up to’ Walter to buy the apple, then he must have been able to do something other than buy the apple.

Compatibilists have responded to consequence argument in two ways. The first way – a direct response – involves arguing that one of the premises of the argument fails. The second way – an indirect response – concedes that causal determinism precludes the ability to do otherwise, but claims that this is not problem for compatibilism. Although there is a vast literature surrounding direct responses to the consequence argument, I won’t be considering this side of the free will debate in this thesis. I will rather be focusing on the literature that has emerged as a result of indirect responses to the consequence argument. John Martin Fischer (1994) defends such a response. He argues that even if causal determinism precludes the ability to do otherwise, it does not preclude the sort control or freedom necessary for moral responsibility.

The appeal to moral responsibility in the free will debate is a longstanding one. Free will is often defined as the ability or power required for agents to be morally responsible for their actions. This isn’t exactly an uncontroversial point, because some philosophers hold that free will is important independent of its connection with moral responsibility or that free will just is

---

6 Humean compatibilists, for example, argue that the fixity of laws of nature is false. On their view, the laws of nature are only ‘set’ at the end of time. See, e.g., Beebee and Mele (2002). However, Russell (1988; 1995) identifies a problem for this Humean compatibilist strategy. He argues there is a tension between both ‘sceptical’ and ‘revisionary’ understandings of Hume’s regularity theory of causation and the argument from compulsion – that is, the argument that to be caused is not necessarily to be compelled. The argument from compulsion entails that an action must have a cause within the agent. However, the regularity theory of causation casts doubt on this: once we accept that there are no necessary connections between events, it is no longer clear that we are able to attribute an action to the agent. This objection is a version of the problem of present luck which is also an objection to libertarian/incompatibilist accounts of free will and moral responsibility.

7 See Harris (2005).

8 This is how McKenna (2012c: 13) defines free will.

9 Steward (2012a, 2012b) is a recent example. Kane (1996) also argues that free will is necessary for more than just moral responsibility, such as creativity and love.
the ability to do otherwise. Neither of these points is clear to me. However, it seems clear to me that linking free will with moral responsibility makes it easier to ascertain what free will amounts to, and avoids any accusation that the free will debate is a mere verbal dispute. The reason it seems that free will requires the ability to do otherwise is because it seems that moral responsibility requires that ability. For instance, when John Locke (1690) proposed a compatibilist-sounding definition of free will, William Molyneux responded that his definition was insufficient for moral responsibility. In other words, Molyneux argued that moral responsibility requires a more robust sense of ‘free will’. So, the only debate about free will worth having is about the sense of free will required for moral responsibility. More importantly, appealing to moral responsibility answers question (2) – it tells us why free will is important.

Fischer’s indirect response to the consequence argument might seem surprising as it seems to be obvious – or part of ‘common sense’ – to many that an individual is morally responsible for an action only if she could have done otherwise than perform that action. In other words, it seems obvious that alternative possibilities are a necessary condition on moral responsibility. Fischer’s response takes its cue from Harry Frankfurt’s (1969) landmark argument against the principle of alternative possibilities (PAP), a principle that Frankfurt takes to enshrine the common sense view that having alternative possibilities is a necessary condition on moral responsibility. Frankfurt argued against PAP by using counter-examples, such as the following:

---

10 Patrick Todd has claimed this in conversation.
11 Chalmers (2011: 530) makes this point, though he also notes that the compatibility debate about moral responsibility might be considered a mere verbal dispute. In §1.4 I endorse (at least for the purposes of this thesis) Pereboom’s (2007: 86) claim that it is the ‘basic desert’ conception of moral responsibility which is at issue in the debate between compatibilists and incompatibilists, thereby avoiding the worry that this debate is a mere verbal dispute.
12 See Harris (2005: 26).
13 Alternative definitions of free will are much older than Locke. The ancient Greeks argued over these matters. The Stoics defended compatibilist versions of free will, whilst Aristotle and Epicurus defended libertarian accounts of free will. For a discussion of this, see Salles (2005).
14 This of course leaves open other senses of ‘free will’ that are unrelated to moral responsibility. Fischer, following in the footsteps of Frankfurt (1971), argues that moral responsibility, and the control or freedom necessary for it, are compatible with the truth of causal determinism. But Fischer is willing to concede for the sake of argument that one sense of free will – regulative control, or the ability to do otherwise – is incompatible with causal determinism.
Suppose someone – Black, let us say – wants Jones to perform a certain action. Black is prepared to go to considerable lengths to get his way, but he prefers to avoid showing his hand unnecessarily. So he waits until Jones is about to make up his mind about what to do, and he does nothing unless it is clear to him (Black is an excellent judge of such things) that Jones is going to decide to do something other than what he wants him to do. If it does become clear that Jones is going to decide to do something else, Black takes effective steps to ensure that Jones decides to do, and that he does do, what he wants him to do. (Frankfurt 1969: 835)

The ‘effective steps’ that Black takes are to ‘manipulate the minute processes of Jones’ brain and nervous system in some … direct way, so that causal forces running in and out of his synapses and along the poor man’s nerves determine that he chooses to act and that he does act in the one way and not in any other’ (Frankfurt 1969: 835-836). This ‘direct way’ might involve Black activating a pre-implanted chip in Jones’ brain that forces Jones’ to act in the manner that Black desires. Suppose that Black wants Jones to kill Walter and that Jones kills Walter without Black having to intervene. Is Jones morally responsible for killing Walter? It seems that he is. Even though Black wants to Jones to do so, Black is no way involved in bringing about Jones’ action. Black is a ‘counterfactual intervener’ which means that he will only compel Jones to kill Walter if it seems that Jones is not going to kill him. However, in his capacity as a counterfactual intervener Black stops Jones from being able to do otherwise, because if Jones showed any sign that he would do otherwise, Black would have intervened and compelled him to kill Walter. Thus, it seems that we continue to judge that an agent is morally responsible even if she lacks alternative possibilities. Hence PAP is false.15

15 Dennett (1984) also argues against PAP. However, he dislikes Frankfurt’s strategy because he is dubious of such ‘intuition pumps’. Dennett instead simply argues that we don’t really care whether an agent could have done otherwise. I think that Dennett’s and Frankfurt’s points are more closely aligned than Dennett realises, though I lack the space to develop this thought here.
Although there is continued debate about Frankfurt’s argument against PAP and the various different counter-examples\textsuperscript{16} that have been proposed to it, there is a growing consensus among many philosophers, as a consequence of Frankfurt’s argument, that alternative possibilities are not a necessary condition on moral responsibility.\textsuperscript{17} This does seem to be a positive result for compatibilists. The worry with causal determinism, after all, is that it apparently precludes alternative possibilities. However, compatibilists can happily concede this because, if Frankfurt’s argument is successful, it seems that moral responsibility (and hence free will) does not require such alternatives. But incompatibilists have not conceded defeat; their arguments against compatibilism have just taken a new form.

What would happen if Black had intervened and compelled Jones to kill Walter? Frankfurt claims that Jones would not be morally responsible for killing Walter. Since Jones would have been compelled by Black to kill Walter, it seems intuitively plausible that Jones would not be morally responsible. Incompatibilists, such as Derk Pereboom (2001), who agree that causal determinism does not preclude moral responsibility by virtue of precluding alternative possibilities, have capitalised on this claim to construct a new argument against compatibilism. According to such incompatibilists, the truth of causal determinism is incompatible with moral responsibility, not because it precludes alternative possibilities, but because it prevents an agent from being the \textit{ultimate source} of her actions. Ultimate sourcehood, they claim, is a necessary condition on moral responsibility. If causal determinism is true, it follows that an individual’s actions will always have a source in events external to the individual and prior to her existence.

\textsuperscript{16} McKenna and Widerker’s (2003) edited collection on the topic summarises the main contours of the debate, though significant progress has occurred since then. Several compatibilists (e.g. Vihvelin 2004; Fara 2008) and several incompatibilists (e.g. Alvarez 2009; Steward 2009) have challenged Frankfurt’s argument against PAP. Both compatibilists (e.g. McKenna 2008b; Fischer 2010) and incompatibilists (e.g. Capes 2012; Shabo 2014) have defended Frankfurt’s argument against these challenges. I won’t discuss these arguments here in what follows.

\textsuperscript{17} Much of the debate follows from the fact that Jones seems to retain some alternative possibilities, despite Black’s presence. However, Fischer (1994) argues that this is a mere ‘flicker of freedom’ and not robust enough to ground judgements of moral responsibility. There is some debate over what a ‘robust’ alternative possibility entails. According to Pereboom (2003: 188), for example, it requires that an agent who is morally responsible for an action ‘could have willed something other than she actually willed such that she understood that by willing it she would thereby have been precluded from the moral responsibility she actually has for the action’. But Pereboom (2001, 2003) argues that such alternatives are \textit{not} necessary for moral responsibility.
So-called ‘source’ incompatibilists claim this would undermine an individual’s moral responsibility for her actions, and that an individual can only be morally responsible for her actions if those actions have an indeterministic source ‘within’ the individual.\textsuperscript{18}

To argue for this claim, source incompatibilists start by imagining cases like the Black/Jones scenario, except they add that Black intervenes. By activating the chip in Jones’ head Black manipulates Jones to kill Walter. Is Jones morally responsible for killing Jones? It seems he is not because the manipulation Jones undergoes compels him to act as he does. If an individual is compelled by another agent, then that individual is arguably under the control of that agent.\textsuperscript{19}

The problem for compatibilists is that incompatibilists can easily design scenarios like this – control cases\textsuperscript{20} – in which an individual is covertly controlled to perform some action (usually something heinous like murder) whilst satisfying the compatibilist conditions on moral responsibility. The problem is that these conditions are causal conditions, and therefore it is conceivable for an individual to be caused to satisfy them by another agent or intentionless force (such as a brain tumour). Thus, it seems that counter-examples to any compatibilist conditions on moral responsibility can be dreamed up. With a counter-example in hand, incompatibilists normally make the further claim that there are no relevant differences between a scenario in which Jones is covertly controlled to kill Walter and a scenario in which Jones (or Jones*) is merely causally determined to kill Walter. If compatibilists cannot explain why the covertly controlled Jones is not morally responsible whilst the merely causally determined Jones is morally responsible, then it seems compatibilists have to concede that the merely causally determined Jones is also not morally responsible; hence compatibilism is false.\textsuperscript{21}

\textsuperscript{18} Of course, indeterminism entails alternative possibilities of some sort, but it does not necessarily entail robust ones.

\textsuperscript{19} Sometimes we might be compelled by an intentionless force. In §5.5, I argue that intentionless forces are capable of controlling things, including individuals.

\textsuperscript{20} In the current literature, these cases are usually called ‘manipulation cases’. I prefer to call them control cases because, as I argue in §5.5, ‘manipulation’ is an imprecise term and doesn’t adequately identify the prima facie responsibility-undermining feature of these cases – namely, that an individual has been covertly controlled. As I note in §5.5, an individual can be manipulated without being covertly controlled, but in such cases the individual is prima facie responsible.

\textsuperscript{21} Another argument for source incompatibilism, which I lack the space to discuss, is the direct argument, which is similar in structure to the indirect argument except argues that moral responsibility is incompatible with determinism.
One thing that is noticeable about control cases, such as the one I have just described, is that they are quite similar to the thought experiments found in the personal identity literature. For instance, imagine that scientists remove part of Clare’s brain and transfer it into Clive’s head (from whom they have already removed the corresponding brain-part). Suppose that the part of Clare’s brain that they remove and implant into Clive is the part that makes Clare a mean bully (this might include the neural substrates of the relevant beliefs, desires, intentions, and the like). When Clive awakes from the operation, he realises that not only does he desire to bully others, but he also believes that he should bully others. This sort of case could be used to argue for a particular claim about personal identity; for example, some might argue that post-operation Clive is a different person to pre-operation Clive because of his change of character. Whether this claim is plausible depends on how we define ‘person’, of course. But suppose post-operation Clive is the same person as pre-operation Clive – that is, post-operation Clive is numerically identical to pre-operation Clive. This case could be used to argue for a particular claim about moral responsibility. Some might argue that since post-operation Clive did not develop or endorse those psychological states which resulted in him becoming a mean bully, that he is not (at least initially) morally responsible for any of this bullying actions.

Contemporary philosophical work on personal identity can be traced back to Locke (1690). Prior to Locke it was usual to think of personal identity as being constituted by sameness of soul. As a preoccupation of the time (which persists into the present day) was to discover how individuals could survive the deaths of their body and enjoy (or suffer) the afterlife, it makes sense that many would appeal to sameness of soul to explain personal identity over time. Souls, after all, are an immaterial substance and so can be imagined to survive the deaths of our (material) bodies. On such a view, we might say that an individual at t₂ is the same person as an individual at t₁ if and only if the individual at t₂ has the same soul as the individual at t₁. Locke

without mentioning alternative possibilities. McKenna (2008c) has argued – convincingly, I think – that the direct argument depends on a control case/manipulation argument to get off the ground. Thus, control cases/manipulation arguments are the crux of the debate between compatibilists and source incompatibilists.
found this sort of view problematic: why would having the same soul as another individual make me the same person as that individual? According to Locke, it is not sameness of substance that matters for personal identity, but rather sameness of *consciousness*.

It is a contentious issue (which I discuss again briefly in §4.5) what exactly what Locke meant by ‘consciousness’, though the most standard interpretation is that Locke meant sameness of memory. So, an individual at $t_2$ is the same person as an individual at $t_1$ if and only if the individual at $t_2$ has the memories of and remembers being the individual at $t_1$. According to Locke, it does not follow that sameness of soul entails sameness of consciousness: an individual $S$ might share a soul with an individual $R$, but $S$ and $R$ might share no other features; hence $S$ and $R$ might have distinct consciousnesses.

Locke’s interest in personal identity was primarily in its connection with *moral responsibility*.\(^{22}\) It is taken to be platitude that an individual can only be morally responsible for her own actions.\(^{23}\) To know which actions belong to which individuals we need an account of personal identity; we need an account that tells us whether an individual at a later time is the same person as an individual at an earlier time if we are properly to praise and blame individuals for their past actions. If personal identity over time were simply a matter of sameness of soul, then it is possible that I share a soul with a serial killer.\(^{24}\) But it wouldn’t be fair to blame me for the serial killer’s actions because we only share the same featureless immaterial substance. This is why Locke thought that personal identity is rather a matter of sameness of consciousness. I am not morally responsible for the serial killer’s actions because I do not have the same consciousness as the serial killer.

---

\(^{22}\) Indeed, according to Strawson, ‘…it looks as if Locke’s theory of personal identity may be completely right, considered specifically as a theory of moral responsibility…’ (2011: 92). Moreover, Strawson argues that Locke intended his account of ‘personal identity’ to be one of moral responsibility (or, rather, what I call an account of diachronic ownership – see §1.4).

\(^{23}\) Cf. Shoemaker (2012). For a list of philosophers who have implicitly or explicitly defended this platitude, see n.79.

\(^{24}\) This scenario is based on a case of Locke’s. He imagines that a friend of his has the same soul as Socrates, but Locke claims that his friend would not be morally responsible for Socrates’ actions.
The similarity between control cases and personal identity thought experiments indicates a link between the free will debate and the personal identity debate. As I noted above, a brain transplant case could either be used to support a claim about moral responsibility or a claim about personal identity. Despite this connection, there has been very little contact between philosophers working in these areas. Most discussions of these topics effectively treat them as distinct and unrelated.

I suspect that the minimal contact between the free will and personal identity debates centres on certain assumption that many philosophers interested in free will implicitly (and sometimes explicitly) make – namely that moral responsibility for past free actions is a simply matter of personal identity. That is, an individual S is morally responsible for a free action $A$ performed at $t_1$ if and only if S is identical to the individual who $A$-ed at $t_1$. As a result, I think that philosophers presume that judgements or intuitions which suggest a diachronic condition on moral responsibility must be suggesting a diachronic condition on being morally responsible at the time of action as opposed to being morally responsible over time. This is a mistake, or so I shall argue.

1.2 Preliminaries

There are some preliminaries to discuss before proceeding. First, I should emphasise that although this thesis is about free will and moral responsibility, I will mainly (as the previous section suggests) talk about moral responsibility. One reason for doing so is that I think we have much clearer intuitive judgements about moral responsibility than we do about free will. It might be that there are numerous things ‘free will’ could plausibly refer to. As I want to focus on the sense of free will required for moral responsibility, I will use the term free will to denote the ability of individuals to exercise control in the strongest manner necessary for moral
responsibility. If we can establish that it is intuitively plausible that an individual is (or is not) morally responsible for \( A \)-ing, we can then infer that she did (or did not) freely \( A \). For example, if we judge that Jones is morally responsible for killing Walter, then we can infer that he freely killed Walter. Conversely, if we judge that Jones is not morally responsible for killing White, then we can infer he did not freely kill Walter. (Of course, as I discuss below, most believe that there are also epistemic conditions on moral responsibility. In all the cases I discuss, we can assume that the epistemic conditions have been satisfied; so we can infer from a judgement about whether an individual is morally responsible to a claim about whether the individual has free will.)

Second, I will only concern myself with moral responsibility for actions and the consequences of actions. Dealing with moral responsibility for omissions and their consequences, and other possible objects of moral responsibility (such as character traits), is far beyond the scope of this thesis. I hope, though, that what I say about actions and consequences will apply to omissions and perhaps also to the other possible objects of moral responsibility.

Third, I will sometimes use the term ‘ownership’ instead of ‘being morally responsible’ – for example, ‘John owns action \( A \)’ instead of ‘John is morally responsible for \( A \)’. The reason, as indicated above, is that there are (at least) two perspectives from which philosophers discuss moral responsibility: the free will debate and the personal identity debate. Philosophers in the free will debate usually aim to provide the conditions that must obtain for an action to belong to an individual in the sense that they at the time at which they perform the action are morally responsible for it. Call these the conditions of synchronic ownership. The compatibility debate is over synchronic ownership. Both sides – that is, compatibilists and incompatibilists – are interested in what makes it the case that an individual is (or is not) morally responsible for an action when she

---

25 To be clear, there are weaker senses of control which are necessary for moral responsibility, such as the control required to be an agent. Cf. McKenna (2012c: 13, n.7).

26 This only applies to direct moral responsibility. In cases of indirect moral responsibility we can infer that the agent is directly morally responsible for something (from whatever they are indirectly morally responsible for) and then we can infer the agent acted freely. I discuss direct and indirect moral responsibility in §2.3.1.
performs it. The other perspective comes from the personal identity debate. Philosophers working from this perspective often aim to provide the conditions that must obtain for an individual at \( t_2 \) to be morally responsible for a free action performed at \( t_1 \). Call these the conditions of *diachronic ownership*. For example, when the accused is standing in the dock facing charges relating to a crime committed many years ago, we need to establish not only that the individual who committed the crime was *at the time* responsible (e.g. they were not sleepwalking or acting under extreme coercion), but also that the individual in the dock bears the right relation to the individual who committed the crime. What is the right relation? The standard answer, of course, is ‘identity’: an agent at \( t_2 \) is morally responsible for a free action performed at \( t_1 \) if and only if the individual at \( t_2 \) is the *same person* as the individual who performed the free action. Thus, the debate between philosophers working from this perspective is normally a debate about the criteria for sameness of person.

As I use the term, ‘synchronic ownership’ is necessary and sufficient for moral responsibility at the time of action. Synchronic ownership has two necessary and jointly sufficient conditions: a freedom condition and an epistemic condition. \(^{27}\) Because I’m only interested in the dispute over free will, I will assume that all the individuals I discuss satisfy the epistemic condition (or conditions). This is a harmless assumption that will allow me to focus discussion on the disagreement between compatibilists and incompatibilists: the freedom/control condition on synchronic ownership (i.e. free will). Synchronic ownership is necessary for diachronic ownership, and diachronic ownership is necessary and sufficient for being morally responsible for some past action at a later time. The purpose of talking about synchronic and diachronic ownership, and not simply being morally responsible, is to make clear

---

\(^{27}\) Although I won’t talk about the epistemic condition or conditions on moral responsibility in what follows, I assume that there is one since almost all philosophers discussing free will and moral responsibility seem to agree that there is an epistemic condition of some sort. Cf. Fischer and Ravizza (1998: 12), Pereboom (2001: 111), Mele (2011), amongst others.
the distinct conditions, and the relationship between those conditions, on being morally responsible at the time of action and being morally responsible at a later time.

I must also distinguish ‘ownership’ from a currently popular term in the moral responsibility literature: attributability. To say an individual owns an action is simply to say that the action belongs to the individual in a sort of way such that she could be held morally responsible for that action (more on this in a moment). But to say an action is attributable has been used to mean that the action belongs to an individual in such a way that she is liable to certain forms of appraisal, such as appraisals of her character. Gary Watson (1996/2004) first introduced the term attributability to distinguish it from the accountability form of moral responsibility. Watson claims that both attributability and accountability are ‘faces’ of moral responsibility. Accountability concerns what we traditionally associate with holding morally responsible – praise, blame, punishment, and reward. I won’t address this distinction in what follows; rather, I will place attributability and accountability under the general title of ‘holding morally responsible’.

Fourth, what do I mean by ‘holding morally responsible’? There are two forms that holding morally responsible can take: private and overt. To privately hold an individual morally responsible is to form a particular attitude about that individual, but not to manifest that attitude in our external behaviour. Suppose Sally privately holds Mike morally responsible for stealing her bike. This means that Sally holds a particular attitude with respect to Mike about a particular (apparent) action of Mike’s. Sally might, for example, now believe Mike is a horrible person or feel angry when she thinks about Mike. But if she only privately holds Mike morally responsible, then her outward behaviour does not change. When she is in Mike’s company, for example, she will treat him as if nothing has happened. Sally will, however, have made a particular judgement about Mike – namely that he is morally responsible for stealing her bike – and she might even

---

28 I borrow this distinction from McKenna (2012c: 26).
feel certain negative emotions towards him, such as resentment, but she does not express those emotions.

To *overtly* hold an individual morally responsible is to form a particular attitude about that individual *and* for that attitude to manifest itself in our behaviour. Sally might, for example, tell Mike he’s a horrible person, she might openly resent him, or she might ban him from her house. Of course, Sally might hold Mike morally responsible, either privately or overtly, in error; it might be that Mike did not actually steal Sally’s bike. If this were the case, then Sally would have *illegitimately* held Mike morally responsible. However, if Mike did actually steal Sally’s bike, Sally would have *legitimately* held Mike morally responsible. Indeed, in certain situations, Sally might legitimately and privately hold Mike morally responsible (e.g. by resenting him), but illegitimately overtly hold him responsible (e.g. by openly resenting him).

Fifth, what is the relationship between ownership and holding morally responsible? Some claim that just because an agent owns $A$ she shouldn’t necessarily be privately or overtly held morally responsible for $A$-ing. For example, Mike might own the action ‘stealing Sally’s bike’ but it might be that Larry cannot legitimately hold Mike morally responsible for that action. Perhaps Larry is also a thief, and so it would be hypocritical of him to hold Mike morally responsible for a similar action.\(^{29}\) I agree that there will be times when an individual owns an action, yet it would be inappropriate for certain individuals to privately or overtly hold her morally responsible. However, my own view is that an individual who is morally responsible for $A$-ing can *in principle* be held morally responsible (either privately or overtly) for $A$. So, when we judge that an individual is morally responsible for $A$-ing, we mean that some *possible* agent can legitimately hold the individual morally responsible for $A$-ing. Conversely, when we judge that an individual is not morally responsible for $A$-ing, we mean that no possible agent can legitimately hold the individual morally responsible for $A$-ing.

\(^{29}\) See Smith (2007). I discuss this more in Chapter 6.
Finally, I should be clear that the concept of moral responsibility at issue in this thesis is the ‘basic desert’ conception. As Pereboom says:

The term “moral responsibility” is used in many ways, but there is one sense that has been at issue in the traditional philosophical debate about free will and determinism. My characterization of it is this: for an agent to be morally responsible for an action is for it to belong to her in such a way that she would deserve blame if she understood that it was morally wrong, and she would deserve credit or perhaps praise if she understood that it was morally exemplary. The desert at issue here is basic in the sense that the agent, to be morally responsible, would deserve the blame or credit just because she has performed the action (given that she understands its moral status), and not by virtue of consequentialist considerations. (2007: 86)

The idea is that an individual deserves praise or blame just because she has knowingly performed a morally exemplary or a morally wrong action; it is in this sense ‘basic’. If an individual deserves praise or blame because of consequentialist considerations – that is, if it is deemed a greater good to praise or blame the individual – then the desert at issue isn’t basic. Indeed, it isn’t clear it should count as desert at all. I will agree with Pereboom that the basic desert sense of moral responsibility is what is at issue in the compatibility debate for the purposes of this thesis.

1.3 Methodology

In this thesis I will be using a wide variety of thought experiments that aim to elicit intuitions or intuitive judgements, which can be used to support arguments for or against a particular philosophical position. There is some controversy over such thought experiments that I must briefly address before continuing. Some argue that our intuitive judgements about thought
experiments are not to be trusted.30 Others argue that our intuitive judgements about outlandish cases are not to be trusted.31 Others – in particular those who engage in experimental philosophy – argue that we ought not to trust any philosopher’s intuitive judgements because these are likely informed by the philosopher’s theoretical commitments, and we should rather discover what the folk think about these matters.32 Despite these worries with thought experiments and intuitive judgements, I will not consider any such objections in what follows. My reason is that my opponents in this thesis all use outlandish thought experiments in order to make their case against the position I am defending – namely, structuralism/compatibilism. So I’m not going to question this implicit methodology because I intend to use it to defend structuralism/compatibilism from those that adhere to such a methodology. There might be some recourse in appealing to the results of experimental philosophy, but I lack the space to engage in such philosophy or with such experimental results.

1.4 Outline

By the end of the thesis, I hope to have shown that manipulation arguments are not a threat to compatibilism. My general (though not my only) strategy is to consider the personal identity issues implicit within the control cases that support manipulation arguments. By showing that these cases are not counter-examples, I thereby aim to undercut the manipulation arguments that these cases are claimed to support. In Part One, I focus mainly on an in-house dispute between compatibilists. In response to certain manipulation arguments, some compatibilists have proposed that synchronic ownership is, in some way, sensitive to an individual’s history. These so-called ‘historicists’ thus claim that there are historical conditions on moral responsibility, and thereby claim to avoid certain manipulation arguments against compatibilism – namely ones that

30 See, for example, Cappelen (2012)
31 See, for example, Wilkes (1988: chapter 1).
32 There is a growing experimental philosophy literature which focuses on the compatibility debate. See, for example, Sripada (2010), Feltz (2012), Murray and Nahmias (2014), Nadelhoffer, Shepard, Nahmias, Sripada, and Ross (2014). For a critique of experimental philosophy see, for example, Finch (2011)
are motivated by hypnosis and brainwashing control cases. However, I believe that the only hope for compatibilism is a structural/non-historical account.

To defend structuralism, I start, in Chapter 2, by showing that historicism has just as counter-intuitive implications as structuralism, and so I conclude that manipulation arguments do not provide decisive motivation for historicism. I then argue that historicism is methodologically inconsistent and once it is made consistent, it collapses into source incompatibilism. Despite defending structuralism by undercutting its compatibilist rival, manipulation arguments unfortunately remain a problem for structuralism.

In Chapter 3, I start my project of defending structuralism from manipulation arguments by sketching the synchronic ownership condition of a new structuralist account, one which I call the ‘structural-narrative view’. I concede that historicists have got one thing right: the ‘standard structuralist conditions’, such as those of Frankfurt (1971), are insufficient for synchronic ownership. Historicists attempt to supplement these conditions with a historical ‘authenticity’ condition; I instead sketch a structural authenticity condition which requires that an individual is able to provide a narrative explanation of an action. This marks my first appeal to narrative views of persons such as Schechtman’s (1996) and Schroer and Schroer’s (forthcoming). I then argue that hypnotised individuals do not satisfy my synchronic ownership conditions, and therefore hypnosis cases do not support a manipulation argument against compatibilism. I then argue that it is probable that brainwashed individuals do not satisfy my conditions either, but I concede that it is conceptually possible that a brainwashed individual does satisfy my conditions.

In Chapter 4, I argue that such brainwashing cases are not counter-examples to my structural-narrative view, once we appreciate its distinctive diachronic ownership condition (or DOC). I argue that the locus of moral responsibility – that is, the entity to which the property of being morally responsible for some action ‘attaches’ – depends upon on the criterion of diachronic ownership. Normally, this has been taken to be personal identity, and so the ‘whole’ temporally extended individual has been taken to be locus of moral responsibility. I then show, following
others, that personal identity is not the DOC. Hence the locus of moral responsibility must not be the ‘whole’ individual. I sketch a new DOC – *narrative connectedness* – and this results in a new locus of moral responsibility, which I call the *narrative self.* This allows me to show that brainwashing cases are not counter-examples to my structural-narrative view, and therefore do not support a manipulation argument against compatibilism.

In Part Two, I turn my attention towards defending compatibilism from the remaining manipulation arguments. In Chapter 5, I focus on clarifying the structure of manipulation arguments. Among other things, I argue, contra Pereboom (2001, 2014), that an inference to the best explanation is not required to support the ‘no difference’ premise of a manipulation argument. I also argue that the ‘no difference’ premise is also unnecessary because manipulation arguments, to show that compatibilism is false, require a counter-example to *all plausible current* compatibilist conditions and such a counter-example (if successful) is sufficient to show that compatibilism is false; I call this argument the *control argument.*

In Chapters 6 and 7, I respond to the remaining control/manipulation arguments – or, rather, the control cases that support those arguments. Depending on the intrinsic features of those cases, I classify them as either threatening or unthreatening. I argue that hard compatibilism – the thesis that covertly controlled and causally determined individuals may be morally responsible for their actions – is the best response to such cases. However, because of the difference between threatening and unthreatening control cases, my response to these types of cases differs. In Chapter 6 I outline my hard compatibilist response to threatening control cases, and in Chapter 7 I outline my hard compatibilist response to unthreatening control cases.
PART ONE

COMPATIBILISM AND PERSONAL IDENTITY
Chapter 2: Against Historicism

2.1 Introduction

Compatibilists hold that moral responsibility is compatible with the truth of causal determinism. Compatibilists also hold that alternative possibilities are not a necessary condition on being morally responsible.\(^{33}\) There are two main forms of compatibilism: structuralism and historicism.\(^{34}\) Structuralists hold that an individual is morally responsible for an action \(A\) only if \(A\) stems from the correct psychological structure at the time of action, though structuralists disagree about what the ‘correct’ psychological structure amounts to.\(^{35}\) Historicians claim that the structuralist conditions are insufficient for moral responsibility; they claim that an individual is morally responsible for an action \(A\) only if she has the right kind of history, but disagree over what the ‘right kind of history’ amounts to.

Given the nature of structuralism, it does seem conceivable that an agent – such as a nefarious neuroscientist – could covertly control an individual while ensuring that the individual satisfies the structuralist conditions on moral responsibility. Many find it intuitive that in such cases the covertly controlled individual is not morally responsible, given the way in which she obtained her psychological structure. When incompatibilists present such cases, they usually

\(^{33}\) At least in the sense that the laws of nature plus the distant past do not determine a unique outcome. Some compatibilists, such as the ‘dispositionalist’ (e.g. Vihvelin 2004, Faro 2008) or classical compatibilists (e.g. Berofsky 2006) hold that causally determined individuals have alternative possibilities in some sense. Humean compatibilists (e.g. Beebe and Mele 2003) reject a necessitarian or governing conception of the laws of nature, and so causally determined individuals might have alternative possibilities in the sense that the laws and the past do not determine a unique outcome. Due to lack of space, I will not discuss these branches of compatibilism in this thesis. Nevertheless, it seems that such compatibilists must also take a stand on the historicism/structuralism debate; so my arguments are relevant to these views.


\(^{35}\) Watson (1999: 363) claims that structuralism should be interpreted as an ‘origin’ or ‘authorship’ theory of moral responsibility, and not as a time-slice one. On this view, the period relevant to moral responsibility is not the time of action but rather around the time of action. But I think that such an account is still a time-slice account of moral responsibility; it just takes a different view on what the relevant time-slice amounts to. I'll set aside this complication in what follows.
assert the further claim that there are no differences relevant to moral responsibility between scenarios involving covert control and scenarios without such covert control in which ‘mere’ causal determinism obtains, and so conclude that compatibilism is false. This argument takes (roughly) the following form:

M1. (Judgement) An individual covertly controlled in manner X to perform action A is not morally responsible for A-ing.

M2. (No difference) There are no differences relevant to moral responsibility between an individual who is covertly controlled in manner X to A and an individual who is causally determined to A.

Therefore,

M3. If causal determinism is true, then no one is morally responsible for anything they do. In other words, compatibilism is false.36

Historicists, however, deny M2. They claim that there is a relevant difference between a covertly controlled and a causally determined individual – namely, the causally determined individual may have satisfied a historical condition on moral responsibility. Given that this argument is supported by an appeal to our intuitive judgements about cases, certain structuralists (as we’ll see) resist the argument by denying that M1 is true. But this, at least initially, seems very counter-intuitive. Thus, the motivation for historicism is that it (a) avoids the counter-intuitive implications of structuralism and (b) avoids (at least) two instances of the above argument for incompatibilism.

36 Formulations of the argument along these lines are found in McKenna (2008a: 143); Mele (2008: 265), Haas (2012: 798), King (2013: 67). Unlike these other formulations, I put the argument in terms of ‘covert control’ rather than ‘manipulation’. This is an important difference, and one that I discuss again in §2.7 and again in Chapter 5. In short, I argue that that individual in control cases seem to lack responsibility because they have been covertly controlled, and not because they have been ‘manipulated’. In §5.5, I argue that it is ambiguous what is meant by ‘manipulation’ with respect to these cases. Nevertheless, to remain in sync with the literature I continue to call this argument-scheme the manipulation argument (or MAS).
I have two goals in this chapter. First, I will argue that historicism has just as counter-intuitive implications as structuralism; therefore, control cases/manipulation arguments do not provide decisive motivation for historicism. In other words, although it might be costly for structuralists to accept such counter-intuitive implications for their view, since historicism has just as counter-intuitive implications compatibilists have no reason to prefer historicism to structuralism. Second, I will argue that historicism is methodologically inconsistent; once it is made consistent, however, it collapses into source incompatibilism. Thus historicism is not a viable option for compatibilists.

This chapter is structured as follows. In §2.2 I outline structuralism, and I argue that control cases are the most devastating objection to structuralism. In §2.3, I argue that structuralists can accommodate the other apparent motivations for historicism. In §2.4-2.6, I argue that historicists must commit to either a psychological or a non-psychological approach to personal identity but that doing so comes with counter-intuitive implications as a consequence of certain thought experiments inspired by those in the personal identity literature. In §2.7, I argue that historicism collapses into source incompatibilism; hence, compatibilists have to be structuralists. This, of course, leaves compatibilists a problem: control cases. It will be the task of Chapters 3-7 to respond to that problem.

2.2 Structuralism

While structuralists agree that there are no historical conditions on moral responsibility, they disagree on what kind of psychological structure is necessary for synchronic ownership. Some, such as Frankfurt (1971) and Watson (1975), claim that an individual is morally responsible (at the time of action) only if there is a certain mesh between her desires or motivations and some other feature of her psychology. For Frankfurt this mesh is between effective first-order desires (those desires which move an individual to action) and second-order volitions (those desires about which first-order desires an individual wants to move her action). According to Frankfurt,
an individual is only morally responsible for actions which stem from effective first-order desires that mesh with her second-order volitions. For Watson this mesh is between motivations (which include desires) and values (an individual’s reflective judgements about what and which courses of action she believes are good). According to Watson, an individual is only morally responsible for actions which stem from motivations that mesh with her values.

Another view is that an individual acts from certain action-producing mechanisms (e.g. practical deliberation, unreflective habit) and she is morally responsible for actions produced by those mechanisms only if those mechanisms are moderately reasons-responsive. This requires that mechanisms are receptive to a stable pattern of reasons to act, which we can know if the individual can recognise that stable pattern. It also requires that the mechanisms are weakly reactive. A mechanism is weakly reactive if in at least one counterfactual sequence of events the mechanism would produce a different action than the mechanism produced in the actual sequence of events. In effect, moderate reasons-responsiveness is a kind of surrogate to an alternative possibilities condition on moral responsibility which also builds in a rationality requirement, i.e. the requirement that the individual must recognise a stable pattern of reasons: rather than the agent requiring the ability to do otherwise in the actual sequence of events, she must do other than she does in the actual sequence in close counterfactual scenarios where her reasons for action differ.  

Mesh accounts and reasons-responsive accounts aim to provide the control condition on synchronic ownership. Let’s call these the ‘standard structuralist accounts’. There are three apparent problems for the standard structuralist accounts: (i) the regress problem; (ii) the weakness of will problem; and (iii) the source/manipulation/control problem. While I won’t respond to (i) and (ii) in what follows, I do think each of the standard structuralist accounts has

---

37 I base this, roughly, on Fischer and Ravizza’s (1998) moderate reasons-responsive theory. Although Fischer and Ravizza also require a historical authenticity condition, there is room for a non-historical reasons-responsive theory. While there are lots of different accounts along these lines, I take the three that I have discussed to be a representative sample.
the resources to resist these problems. My strategy in what follows is simply to show that (i) and (ii) are not a problem for all standard structuralist accounts, and thus are not problems for structuralism in general. To respond to versions of (iii), however, I argue in Chapter 3 that the standard structuralist account must be supplemented with a structural authenticity condition.

The regress problem seems to only be a problem for Frankfurt. On his view, synchronic ownership requires a mesh between effective first-order desires and second-order volitions. But if it is possible to form second-order volitions, it is in principle possible to form third-order volitions, fourth-order volitions, and so on ad infinitum. Thus an infinite regress looms. Aware of this problem, Frankfurt (1971) invokes the concept of ‘identification’ to avoid the regress problem. He claims that:

It is possible … to terminate such a series of acts without cutting it off arbitrarily.

When a person identifies himself decisively with one of his first-order desires, this commitment “resounds” throughout the potentially endless array of higher order desires. (Frankfurt 1971/1988: 21)

Identification is here characterised as a ‘resonance’ condition that stops the potential need to form ever higher-order volitions. Watson (1975: 218-219), however, complains that Frankfurt’s solution is unsatisfactory because identification is arbitrary. Watson claims that his account avoids such a regress because there is no hierarchy of values. A reasons-responsive account avoids this problem for a similar reason. Given this, the regress problem is not devastating for structuralism in general.

Dealing with cases of weakness of will (or akrasia) seems to be a problem for mesh theories. According to many philosophers (e.g. Mele 1995; Fischer and Ravizza 1998: 42-43;

---

38 Frankfurt (1988: 167-169) offers a further reply to Watson, but I lack the space to discuss this response here. Moreover, I think that the way to solve this problem is to posit a narrative condition on synchronic ownership, as I do in Chapter 3. I also lack the space to discuss this potential response here, though.
Pereboom 2001: 106), we hold individuals morally responsible for acting from weakness of the will. For example, suppose an individual A-ing does not express one of her values or the effective first-order desire which led her to A does not mesh with a higher-order volition of hers. She acts weakly in that she acts against what she judges is the best thing to do, or because she acts in a way she doesn’t want to want to act. It is allegedly intuitively plausible that we might still hold the individual morally responsible for A-ing. According to a reasons-responsive view, though, an individual may still be morally responsible because, while she acts weakly, there are close counterfactual scenarios in which she wouldn’t have acted this way; hence she is still responsive to reasons. Thus, such cases might motivate a reasons-responsive view, because it seems that mesh theories do not have the resources to avoid this problem. I think there is room for mesh theorists to respond to this problem, but I won’t consider any responses here. The fact that a reasons-responsive account avoids the weakness of will problem is sufficient to show that the weakness of will problem is not devastating for structuralism.

This takes us to the third and most important challenge to structuralism: the source/manipulation/control problem. It seems that an agent – perhaps a nefarious neuroscientist – could covertly control an individual to perform some heinous action while ensuring that the individual satisfies the standard structural conditions on synchronic ownership. Consider first a case without covert control:

_Evil Chuck._ Chuck enjoys killing people, and he ‘is wholeheartedly behind’ his murderous desires, which are ‘well integrated into his general psychic condition’ (Frankfurt 2002: 27). When he kills, he does so ‘because he wants to do it’ (Frankfurt 2002: 27), and ‘he identifies himself with the springs of his action’ (Frankfurt 1988: 54). When he was much younger, Chuck enjoyed torturing animals, but he was not

---

39 One response involves claiming that there are different types of responsibility, and then conceding that individuals cannot be responsible in the basic desert sense for weak willed actions, but claiming that they can be responsible in some other sense for such actions.
wholeheartedly behind this. These activities sometimes caused him to feel guilty, he experienced bouts of squeamishness, and he occasionally considered abandoning animal torture. However, Chuck valued being the sort of person who does as he pleases and who unambivalently rejects conventional morality as a system designed for and by weaklings. He freely set out to ensure that he would be wholeheartedly behind his torturing of animals and related activities, including his merciless bullying of vulnerable people, and he was morally responsible for so doing. One strand of his strategy was to perform cruel actions with increased frequency in order to harden himself against feelings of guilt and squeamishness and eventually to extinguish the source of those feelings. Chuck strove to ensure that his psyche left no room for mercy. His strategy worked. (Mele 2013a: 169)

This sort of case is then compared with a case in which an individual is brainwashed to have the same ‘system of values’ as Chuck. Mele provides the following case:

*Brainwashed Beth.* When Beth crawled into bed last night she was an exceptionally sweet person, as she always had been. Beth’s character was such that intentionally doing anyone serious bodily harm definitely was not an option for her: her character – or collection of values – left no place for a desire to do such a thing to take root. Moreover, she was morally responsible, at least to a significant extent, for having the character she had. But Beth awakes with a desire to stalk and kill a neighbor, George. Although she had always found George unpleasant, she is very surprised by this desire. What happened is that, while Beth slept, a team of psychologists that had discovered the system of values that make Chuck tick implanted those values in Beth after erasing hers. They did this while leaving her memory intact, which helps account for her surprise. Beth reflects on her new desire. Among other things, she
judges, rightly, that it is utterly in line with her system of values. She also judges that she finally sees the light about morality – that it is a system designed for and by weaklings. Upon reflection, Beth ‘has no reservations about’ her desire to kill George and ‘is wholeheartedly behind it’ (Frankfurt 2002: 27). Furthermore, the desire is ‘well integrated into [her] general psychic condition’ (Frankfurt 2002: 27). Seeing absolutely no reason not to stalk and kill George, provided that she can get away with it, Beth devises a plan for killing him, and she executes it – and him – that afternoon. That she sees no reason not to do this is utterly predictable, given the content of the values that ultimately ground her reflection. Beth ‘identifies [herself] with the springs of her action’ (Frankfurt 1988: 54), and she kills George ‘because [she] wants to do it’ (Frankfurt 2002: 27). If Beth was able to do otherwise in the circumstances than attempt to kill George only if she was able to show mercy, then, because her new system of values left no room for mercy, she was not able to do otherwise than attempt to kill George. (2013a: 169-170)

Since it is (allegedly) intuitively plausible that Beth isn’t morally responsible for killing George while Chuck is morally responsible for his evil actions, it seems there must be a relevant difference between Beth and Chuck. Mele concludes that the difference between Chuck and Beth must be historical, because they are psychological twins (in the respects relevant to moral responsibility), and so there must be a historical condition on synchronic ownership. This case appears to be a counter-example to not only Frankfurt’s structuralism but all the standard structuralist accounts, since whatever features of an agent’s psychological structure that

---

40 Note that in these sorts of cases the fact that Beth is covertly controlled by other agents is alleged to be unimportant. Mele claims that we can replace the agent-manipulators with an intentionless force, such as a brain tumour, and still elicit the non-responsibility judgement (2009a: 168, n. 11).

41 Note that since this debate is between compatibilists it can be taken as uncontroversial that Chuck is morally responsible. Incompatibilists thus ought to set their intuitions to ‘compatibilist’ for the remainder of Part One.
structuralists might point to as being relevant to the individual’s synchronic ownership can presumably be met in Brainwashed Beth. Hence this case appears to motivate historicism.  

Frankfurt, however, disagrees and takes a notoriously hard line in response to these sorts of cases. He writes:

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. It is irrelevant whether those causes are operating by virtue of the natural forces that shape our environment or whether they operate through the deliberative manipulative designs of other human agents. We are the sorts of persons we are; and it is what we are, rather than the history of our development, that counts. The fact that someone is a pig warrants treating him like a pig, unless there is reason to believe that in some important way he is a pig against his will and is not acting as he would really prefer to act. (2002: 28)

Frankfurt seems to bite the bullet by claiming (in effect) that individuals who satisfy his conditions on moral responsibility (i.e. synchronic ownership) are morally responsible no matter how they came to satisfy those conditions. While Frankfurt is usually considered to be biting the bullet in response to counter-examples to his conditions on synchronic ownership, I think there

---

42 Fischer and Ravizza (1998: 195-196) argue that so-called ‘tracing cases’ also motivate historicism. I consider such cases in §2.3.1.
is much more to his reply. Indeed, the synchronic ownership condition, which I sketch in §3.2, is inspired by Frankfurt’s implicitly proposed ‘character’ condition.

Historicists have been less than impressed with Frankfurt’s response. Mele, for example, says that ‘if compatibilists were to have nothing more attractive to offer than Frankfurt’s ahistorical view of moral responsibility and freedom, compatibilism would be in dire straits’ (2003: 294). And Fischer voices similar concerns when he says that ‘The moral I draw from [control cases] is that an adequate theory of moral responsibility will attend to the history of an action, and not simply to its current time-slice characteristics’ (1994: 209). Although historicists reject Frankfurt’s bullet-biting response, I want to defend it and, in doing so, structuralism (and, as we’ll see, compatibilism). The first part of my strategy for doing so will be to show that historicists have to bite equally unpalatable bullets. Before that, I will argue that control cases are the only clear motivation for historicism.

### 2.3 Other potential motivations for historicism

Apart from control cases, there are various other proposed motivations for historicism. In this section, I argue that structuralists can accommodate each of them; hence only control cases motivate historicism.

#### 2.3.1 Tracing Cases

Normally we must act freely in order to be morally responsible for our actions. But there are some cases where we seem to not act freely – at least at the time of action – and yet we are still intuitively morally responsible for our actions. Intoxication cases are often discussed in this context. When I’m blind drunk, I might not act freely, but that doesn’t mean I’m not morally responsible for my actions; otherwise, people would just get drunk to avoid responsibility for their actions. As long as I got drunk of my own free will (and satisfied certain epistemic
conditions), then I am morally responsible for my drunken actions even if those actions are not free actions.

Fischer and Ravizza (1998: 49-50) appeal to such ‘tracing cases’ in making their case for historicism. They claim that such cases are further evidence that moral responsibility is historical in some sense. The drunken person, on this view, is derivatively (or indirectly) morally responsible for her action because it traces back to or derives from an earlier action of hers which she is directly morally responsible for. Such cases might, then, seem to be a problem for structuralism, because structuralists claim that moral responsibility for actions depends only on facts about the agent’s psychology at the time of action.

I struggle to see why this sort of case is a problem for structuralism. As far as I can see, an individual is morally responsible for her drunken behaviour if it is a consequence of an earlier free action (i.e., getting drunk). To be morally responsible for a consequence $C$ of an action $A$, an individual must know (or should know) that $C$ is a potential consequence of $A$ and she must $A$ freely. Likewise, to be morally responsible for a drunken action $D$, an individual must know (or should know) that she could $D$ as a result of getting drunk and she must get drunk freely. Thus, an individual is morally responsible for $D$-ing in the same sort of way that an individual is morally responsible for $C$. An individual is responsible for consequences and drunken actions if they can be traced back to an earlier free action of the individual’s. I propose, then, that drunken actions should simply be considered consequences of earlier free actions. Since structuralists have no problem accepting that an individual is morally responsible for consequences of her actions, they have no problem accepting that an individual is morally responsible for her drunken actions. Hence, intoxication tracing cases do not pose a problem for structuralism.

---

44 See also McKenna (2012a: 166-167).
Other sorts of tracing cases might seem more problematic. These are more ‘long-term’ than intoxication cases.\(^{45}\) Suppose Alice has a character such that she punches people in the face when they annoy her. One day Bill annoys her, so Alice punches him in the face. Is Alice morally responsible for punching Bill? Proponents of tracing, such as Kane (1996) and Fischer and Tognazzini (2009), claim that this depends on whether Alice freely set her character this way. It seems implausible for structuralists to claim that Alice is morally responsible for punching Bill in the same way that an individual is morally responsible for a consequence of her earlier free actions. The plausibility of this move stems from the fact that intoxicated actions hardly seem like actions.\(^{46}\) Alice’s punching of Bill, though, certainly seems like an action. So structuralists cannot endorse the same response to long-term tracing cases.

But structuralists have a much more straightforward response: Alice is morally responsible for punching Bill only if she has the right kind of psychological structure at the time of action. In other words, the structuralist conditions are sensitive to what count as a proper or an improper ‘set’ character. Notice that it seems that whether we judge that Alice is morally responsible depends on her psychology at the time of action. If she were (for example) extremely mentally ill, we would not judge that she is morally responsible for punching Bill. This seems true even if Alice is mentally ill of her own doing. Even those who appeal to tracing cannot plausibly deny this; otherwise, they would have to accept that those who are extremely mentally ill might be morally responsible for their actions, and that’s a bad result for any account of

\(^{45}\) I draw here from the debate between Vargas (2005), and Fischer and Tognazzini (2009). These cases are also discussed by Kane (1996).

\(^{46}\) Note that the sort of intoxicated actions at issue here are those of an \textit{extremely} intoxicated individual. Individuals who are only slightly to moderately drunk plausibly still act freely. For intoxication cases to motivate a tracing component in an account of moral responsibility, the individual must be so intoxicated that it is implausible that she still acts freely. In short, intoxication cases are much more complicated than certain philosophers portray them to be. I suspect garden-variety cases of intoxication would not motivate a tracing component or need to be explained in the manner I have suggested. The cases that might motivate a tracing component are genuine cases; it’s just they are much more rare (and perhaps even only conceptual possibilities) than proponents of tracing have admitted.
synchronic ownership. Hence, structuralists get the right results in cases like these, without needing to appeal to tracing. Therefore tracing cases do not motivate historicism.47

2.3.2 The development of moral agents

It is a fact that it takes some sort of process to turn a mere human being into a moral agent. Fischer and Ravizza (1998: 208-210) provide one story that might accurately represent the process that human beings must go through in order to become moral agents.48 Does this provide historicists with motivation? No. Structuralists can happily accept that normally a process is required to turn a mere human being into a moral agent. All that a structuralist is committed to is that this process is in principle dispensable. Structuralists should instead say that the process itself is not important; rather, the outcome of that process is. In other words, an individual is morally responsible for an action only if it issues from the right kind of psychological structure no matter how that structure came to be. After all, it is the outcome of that process which is important, and not the process itself.

Here’s an analogy. To make cake, we normally go through a specific process. First, we get some butter, some sugar, some eggs, and some flour. We then mix the ingredients, put the batter into a suitable container, and then place the container in an oven at a certain temperature for a certain amount of time. Finally, we remove the container from the oven. The result should be some (hopefully tasty) cake. Now it is conceptually possible that scientists, or an extremely powerful being, could bring some cake in to existence in an instant. Suppose this cake is qualitatively identical to the home-made cake. Despite being instantly made, the instant-cake is just as much a cake as the home-made cake. Structuralists, I claim, should think of moral responsibility as analogous to cake. Normally it takes a specific process to turn ingredients into a

---

47 The arguments here are admittedly brief, though they are sufficient for my current purposes. I plan to develop them in future work. Indeed, I think that tracing is dubious and that only structuralists can actually provide a plausible explanation of our judgements in these cases and other cases which are problematic for proponents of tracing.

48 Haji (1998: 108-123) argues that this is a motivation for historicism.
cake, but this process can be dispensed with in principle. Something’s being cake is not constituted, even in part, by its being created through this mixing and baking process.

2.3.3 The circumstances of development

The final potential motivation for historicism that I will discuss is that our judgements of moral responsibility seem to be affected by learning about an agent’s history. Although we might initially judge that an agent is morally responsible for $A$-ing, this judgement often changes once we learn about the agent’s traumatic past that likely led to them $A$-ing. Gary Watson (1987/2004), for example, considers the real life story of Robert Harris. Watson quotes from a newspaper article (Corwin 1982) which first describes in great detail the heinous actions that Harris performed. We get the clear idea that Harris was a cruel and sadistic person, and that he thought about his actions before he performed them. Harris thus seems like someone who is morally responsible for his actions. However, the rest of the article seems to change our judgement. We learn about the full details of Harris’s troublesome past: his abuse from his father, his early-life incarceration and the sexual violence he endured then. Learning about Harris’s past seems to affect our judgement about his moral responsibility; it seems to at least diminish the extent to which we think he is morally responsible, and it might even render him not morally responsible at all. Given that our judgements about moral responsibility are affected by history, this seems to provide some evidence that moral responsibility (qua synchronic ownership) is a historical in some sense.

However, this real-life case doesn’t motivate historicism at all. Structuralists can explain why our judgements about moral responsibility change here without conceding anything significant to the historicist. All a structuralist need say is that learning about an agent’s history is one way that we learn about an agent’s psychology at the time of action. We do this all the time with new friends. Although we might initially judge that a friend is very confident, after we learn that she was bullied as a teenager, we might judge that the confidence she displays is a front to
cover the insecurities she still has. It doesn’t seem much of a stretch to think that this might cover whether other individuals satisfy the structuralist conditions on synchronic ownership.

Normally we use an individual’s behaviour to learn what her values, desires, and reasons are, but we could easily use facts about her history to learn this too. In this way, structuralists can claim that an individual’s history is epistemically relevant to her moral responsibility, because it is one way of learning whether or not she is morally responsible for an action (by learning what her psychology is like at the moment of action). This does not affect the structuralist’s core claim that only factors at the time of action are constitutively relevant to synchronic ownership.49

Control cases are therefore the only motivation for historicism. In the next few sections, I argue that historicism has counter-intuitive implications that structuralism does not as a result of thought experiments inspired by those in the personal identity literature.

2.4 Historicism and personal identity

There is a prima facie case to be made that Brainwashed Beth is not a counter-example to structuralism because post-manipulation Beth (post-Beth) is, in fact, a different person to pre-manipulation Beth (pre-Beth). While this response does not succeed as it stands, as we shall see, it serves to illuminate the relevance of considerations of personal identity to the debate between historicists and structuralists, and hence sets the scene for the (better) arguments to come.

Notice that the Brainwashed Beth case resembles the sorts of case discussed in the personal identity literature. Consider the following example from Locke:

Could we suppose two distinct incommunicable consciousnesses acting the same body, the one constantly by day, the other by night; … I ask … whether the day and

49 According to Fischer and Ravizza (1998: 188, n.36), Frankfurt also proposed this sort of strategy in conversation. Vargas (2013: 269) also suggests that one way we learn about an agent’s psychology at the time of action is by considering her history.
the night man would not be two as distinct persons as Socrates and Plato?

(1690/1975: 48)

In Locke’s case, there is one body with two distinct consciousnesses: one that operates during the day, and one that operates during the night. Locke thinks that it is possible that there can be more than one person in a human body. In this case, we would judge that the day-person would be morally responsible for his and only his actions, whilst the night-person would be morally responsible for his and only his actions, because they are numerically distinct. As such, the day-person’s responsibility for a particular action would not transmit to (or be shared with) the night-person, because the action in question would not be one of the night-person’s actions.50

Of course, this case differs somewhat from the Brainwashed Beth case. Pre- and post-Beth are not ‘distinct incommunicable consciousnesses’ because post-Beth retains the memories of pre-Beth. Consider, then, a modern rendition of Locke’s case:

From an early age Leland has occasionally been possessed by a spirit named ‘Bob’.

When Bob possesses Leland this involves Bob taking control of Leland’s body, and suppressing Leland’s agency. Bob then uses Leland’s body to perform many monstrous acts – though Leland remains conscious of what is happening throughout.

Both Leland and Bob satisfy the structuralist compatibilist conditions on synchronic ownership, though Leland only does so when he isn’t possessed by Bob. And Leland does not endorse Bob’s actions even though Leland feels as though he performed Bob’s actions.51

50 Although Locke talks of the ‘day man’ and the ‘night man’, he presumably means ‘person’ by this use of ‘man’, since he distinguishes between the man (the body or biological organism) and the person (the locus of moral responsibility). Hence I have instead talked in terms of the day-person and the night-person.

51 This case comes from the television programme Twin Peaks.
Is Bob morally responsible for performing those heinous acts? It seems clear that he is. It’s also clear that Leland isn’t morally responsible for those acts. It would, of course, be unjustified to blame Leland for Bob’s actions, but that doesn’t mean that Bob isn’t responsible for his actions. This case is in many ways parallel to the pre-Beth/post-Beth scenario. Pre-Beth is just like Leland, and post-Beth is just like Bob. The important question is whether post-Beth is morally responsible for her actions. If post-Beth is numerically distinct from pre-Beth, then, just as Bob is responsible for his actions, it seems that post-Beth can be responsible for her actions without pre-Beth being morally responsible for those same actions.

Suppose that pre- and post-Beth are indeed numerically different persons. How would this undermine the case for historicism? Well, I suggest that one motivation for claiming that Beth is not morally responsible for murdering George is the thought that she is an ‘exceptionally sweet’ person who has been covertly controlled into committing a heinous act (via manipulation of her values). The reason we are inclined not to attribute moral responsibility to post-Beth is because we believe that she is pre-Beth; we believe that pre-Beth has simply been changed against her will and so it would be preposterous to think she is morally responsible for actions that result from the changes that have been made to her. However, it is pre-Beth who was exceptionally sweet, and she, unfortunately, no longer exists. Post-Beth, by contrast, is not and has never, in the short time she has existed, been anything other than a moral monster. So we can happily say that post-Beth is morally responsible for murdering George while denying that the exceptionally sweet pre-Beth is morally responsible. In other words, just because pre-Beth is not morally responsible, it doesn’t follow that post-Beth is not morally responsible.

It seems, then, that structuralists can account for our unwillingness to attribute moral responsibility for murdering George to Beth by distinguishing pre-Beth, who lacks responsibility, from post-Beth, who is morally responsible, without appealing to any historical conditions on

---

52 Some might think that, even if post-Beth is numerically distinct from pre-Beth, the mere fact that post-Beth has been covertly controlled undermines her moral responsibility. I set aside this possibility until §2.7 where I argue that this possibility is a problem for historicists, and I respond to this point in §4.8.
moral responsibility. Hence if pre- and post-Beth are numerically distinct, then Mele’s case fails to demonstrate that structuralism fails; therefore historicism is without motivation.

Unfortunately, however, Mele provides a response to the suggestion that pre- and post-Beth are different persons:

The [Beth] case … might have prompted worries about personal identity. Is the transformed [Beth] the same person as the pre-transformation [Beth]? Is Beth, after becoming a [Chuck] ‘twin’, the same person as the earlier Beth? This is not the place to advance a theory of personal identity. But, surely, the pre- and post-transformation agents have much in common? [Beth] just before [her] transformation ([pre-Beth]) is much more similar, on the whole, to [Beth] just after it ([post-Beth]) than [she] is to neonate [Beth] or toddler [Beth]. Still [pre-Beth] is the same person as the neonate and toddler [Beth], in a familiar ‘personal identity’ sense of ‘same person.’ So what is to prevent [her] from being the same person, in the same sense, as [post-Beth]? It is worth noting, further, that [pre-Beth] and [post-Beth] may be strongly psychologically connected, in Parfit’s sense (1984: 206). They may be such that the number of direct psychological connections between them ‘is at least half the number that hold, over every day, in the lives of nearly every actual person.’ (1995: 175, n.22)

It is unclear what Mele means by the claim that post-Beth has ‘more in common’ with pre-Beth than pre-Beth has with neonatal-Beth or toddler-Beth. When it comes to personal identity we have, broadly speaking, a choice between ‘physical’ or ‘bodily’ or ‘biological’ (i.e. substance) accounts on the one hand and psychological accounts on the other. Thus, we can elide the

---

53 In this and later quotations Mele is actually talking about different characters. The cases in which those characters featured are structurally similar, so Mele’s point applies here. Indeed, Mele often refers to this point in later works to ward off any worries about personal identity – for example, Mele (2009a: 176; 2009b: 465; 2013a: 170).
question whether Mele thinks the ‘more in common’ is a matter of bodily similarity or psychological similarity, and simply grant that on a physical/bodily/biological account, pre-Beth and post-Beth are the same person. But Mele also implicitly claims that pre- and post-Beth are the same person according to a psychological account. According to Parfit (1984), the holding of at least half the number of direct psychological connections that hold over every day in the lives of nearly every actual person constitutes a criterion of personal identity. So if Mele is right about the direct psychological connections between pre-Beth and post-Beth, it follows, according to Parfit’s criterion, that they are the same person.

It seems, then, that whichever account of personal identity the historicist endorses, post-Beth is the same person as pre-Beth; this undermines the anti-historicist argument above. If pre- and post-Beth are the same person after all, then our unwillingness to attribute moral responsibility to Beth cannot be explained by distinguishing pre-Beth (who lacks responsibility) from post-Beth (who doesn’t).

The argument that Brainwashed Beth is not a counter-example to structuralism therefore fails. However, it serves to illuminate the connection between historicism on the one hand and facts about personal identity on the other: as we have just seen, Mele’s claim that pre-Beth and post-Beth are the same person is crucial to the success of the Brainwashed Beth case as a counter-example to structuralism. In the following two sections I argue on the basis of thought experiments inspired by those in the personal identity literature that historicism has just as counter-intuitive implications as structuralism has, and that this undermines the motivation for historicism that control cases are supposed to provide.

---

54 In fact, the psychological criterion also requires a nonbranching clause because psychological continuity does not necessarily hold one-one; but I shall ignore this clause since branching is not at issue in the Beth case. I discuss theories of personal identity in more detail in Chapter 4.
2.5 Historicism and the psychological approach

In this section, I will argue that historicists are committed to a non-psychological approach to personal identity. Consider the following case:

*Evil Chuck, Part Two*: Neuroscientists decide to brainwash Chuck. When the neuroscientists brainwashed Beth they implanted Chuck’s system of values, and left everything else in place (call this her ‘background psychological contents’). When brainwashing Chuck, the neuroscientists decide to leave his system of values in place whilst replacing his background psychological contents. The neuroscientists replace Chuck’s background psychological contents with ones which are qualitatively similar to Beth’s. Post-brainwashing Chuck’s first overt action is to murder his neighbour.

Is Chuck morally responsible for the murder? It seems intuitively plausible that he is. After all, he was morally responsible when he had one set of memories (and other background contents), and it doesn’t seem to make a difference to his moral responsibility that he now has different memories (and other background contents), because the psychological contents that led to his murderous actions – that is, his system of values – are present post-brainwashing. Of course, that’s assuming that post-brainwashing Chuck’s (post-Chuck’s) new psychological contents integrate with pre-brainwashing Chuck’s (pre-Chuck’s) system of values. If they don’t, then post-Chuck would not be a coherent agent, and so would not be morally responsible for his action. But historicists cannot question the assumption that post-Chuck is coherent; if they did then structuralists could equally claim that Brainwashed Beth is not a counter-example. Post-Chuck and post-Beth are, after all, psychologically identical: they have the same background psychological contents and the same system of values. So if it were true that post-Chuck is not a coherent agent, the same would be true of post-Beth, and vice versa. Hence this response is off the table for the historicist. If it’s intuitively plausible that post-Chuck is morally responsible,
then we need to know whether post-Chuck has satisfied the relevant historical condition on moral responsibility. This seems to require that post-Chuck has a past (more on this in a moment); thus, it matters whether or not pre- and post-Chuck are the same person.

But is post-Chuck the same person as pre-Chuck? That, of course, depends on which account of personal identity is true. Let’s return to Mele’s claim that pre- and post-Beth are strongly psychologically connected – that is, that they are such that the number of direct psychological connections between them is at least half the number that hold, over every day, in the lives of nearly every actual person (call the total number of connections $N$) – and hence they are the same person. If Mele is right about Beth, it follows that pre-Chuck and post-Chuck are different persons, according to a psychological approach to personal identity. After all, if the number of direct psychological connections that hold between pre- and post-Beth is at least $N/2$, then the relevant systems of values (Beth’s and Chuck’s) must constitute less than $N/2$. But since his system of values is all that Chuck retains post-brainwashing – all the rest of his psychological contents are like Beth’s – post-Chuck is not strongly psychologically connected to pre-Chuck.

So, given Mele’s assumption about Beth, which is what allows him to dodge the defence of structuralism given in §2.4, pre-Chuck and post-Chuck are different persons (according to the psychological approach, that is). Despite this, it still seems intuitively plausible that post-Chuck is morally responsible for the murder. This presents a problem for the historicist because she requires a morally responsible individual to have a history which is hers; and facts about which history is the individual’s depend upon facts about personal identity. An individual $A$’s history is the period in which there is an individual who is numerically identical to $A$.

Historicists, however, might disagree here. According to Fischer and Ravizza’s (1998) positive historical condition on moral responsibility, an individual does indeed require a history – specifically, they argue that an individual must have ‘taken responsibility’ for her action-producing mechanisms. Taking responsibility requires that an individual see herself as the source
of her actions, recognise that she is an apt target for the reactive attitudes, and her view of herself be based appropriately on the evidence (1998: 210-213). But other historicists, such as Mele (1995, 2006) and Haji (1998, 2013), defend a negative historical condition on moral responsibility. This requires than an individual lack a history of a particular sort – namely, one that includes events which ‘bypass’ an individual’s control over her mental life. As such, Mele and Haji hold that moral responsibility does not require that an individual have a past; they only require that individuals who have pasts have a particular sort of past. However, negative historicism is unmotivated: the only plausible historical condition is a positive one, as I will now argue.

Although Mele, for example, claims to advocate a negative historical condition, in making his case for this condition he appears to be appealing to a positive historical condition. Consider the following passage:

The salient difference between [Chuck] and Beth is that [Chuck’s] practically unsheddable values were acquired under [his] own steam, whereas Beth’s were imposed upon her. [Chuck] autonomously developed [his] values (we are entitled to suppose); Beth plainly did not. (Mele 1995: 155; my emphasis)

Practically unsheddable values are those values that an individual cannot modify or attenuate over a particular period; they are a deeply entrenched part of a person (Mele 1995: 153). According to Mele, when an individual is manipulated to have such values, they are not, at least initially, morally responsible for the resulting actions. It is only when those values were endorsed or acquired ‘under an individual’s own steam’ – that is, without interference from external forces – that an individual can then be morally responsible for actions that stem from such values. This

---

55 I discuss taking responsibility in more detail in §3.2.
certainly sounds like a positive historical condition on moral responsibility: a condition that requires the individual to have a past, during which time he has acquired his system of values ‘under his own steam’.

The point here is that, while Mele officially claims only to be offering a negative historical condition on moral responsibility, in the course of trying to persuade us that Chuck is, but Beth is not, morally responsible he illicitly appeals to a positive condition: that of having acquired one’s unsheddable values under one’s own steam. Indeed, it seems that any plausibility that historicism has is as a result of this implicit historical condition. After all, the reason we are supposed to think that post-Beth is not morally responsible is because she acts from unsheddable values that she has not developed or endorsed under her own steam, whilst Chuck is morally responsible because he has done so. Hence, historicists have no option but to endorse a positive historical condition of some sort. In other words, historicists are committed to the claim that a morally responsible individual – at least one who acts from unsheddable values – requires a past.\(^{56}\)

This means that if post-Chuck is a different person to pre-Chuck (as the psychological approach says), then he won’t be morally responsible because he comes into existence as a fully developed individual with ‘unsheddable values’ – that is, values which an individual cannot practically get rid of. And this will undermine his responsibility, because it isn’t post-Chuck who developed such values. The upshot is that the historicist cannot account for post-Chuck’s moral responsibility for the murder if she endorses a psychological approach to personal identity.

The obvious move for the historicist to make at this point is to embrace a non-psychological account of personal identity. The most plausible alternative they could endorse is Eric Olson’s biological approach (1999),\(^{57}\) according to which persons are essentially human

---

\(^{56}\) For further support of the claim that historical conditions must have a positive strand, see McKenna (2012a: 167-169).

\(^{57}\) I will focus upon the biological approach, and in particular Olson’s (1999) biological approach, for the sake of simplicity. These points will apply to other substance-based accounts of personal identity, such as Thomson’s (1997) bodily continuity theory and Swinburne’s (1984) dualist account.
animals; hence the persistence conditions of persons over time are just the persistence conditions of human animals. Thus radical breaks in psychological continuity do not result in a change of person: pre- and post-Beth are uncontroversially the same person (since they are the same human animal and are both persons), as are pre- and post-Chuck. Thus, by embracing the biological approach (or indeed any non-psychological approach), the historicist can make sense of post-Chuck’s moral responsibility: post-Chuck’s history extends backwards in time, throughout the intervention of the neuroscientists, encompassing his very conscious and deliberate formation of his set of values. (Of course, Chuck, post-manipulation, can’t remember any of this; he woke up this morning with memories similar to Beth’s. Nonetheless, it was he himself who developed those values.)

Of course, the other alternative for the historicist would be simply to bite the bullet when it comes to Evil Chuck Part Two – holding that post-Chuck is not morally responsible for the murder – and continue to endorse a psychological account of personal identity. But not attributing moral responsibility to post-Chuck is every bit as implausible as judging Beth to be morally responsible. Thus the historicist who took this option would be committed to a position that is no more plausible than the Brainwashed Beth case (allegedly) shows structuralism to be. And this would undermine the historicist’s complaint that Frankfurtian bullet-biting in the case of Brainwashed Beth is unacceptable, since the historicist would now be guilty of biting an equally unpalatable bullet.

2.6 Historicism and the biological approach
As long as historicists endorse a non-psychological approach to personal identity, it seems they can accommodate our judgements in cases like Brainwashed Beth and Evil Chuck Part Two. However, a further thought experiment will show that this option also spells trouble for the historicist. Consider the following case:
Cerebrum Transfer: One summer’s evening, Chuck decides to kill an innocent family. Before he can, a team of maverick neurosurgeons remove Chuck’s cerebrum (the psychology-conferring part of his brain). The neurosurgeons then transfer Chuck’s cerebrum into another body, which they have just constructed out of the appropriate organic materials. The neurosurgeons name the resulting person ‘Chuckie’. Chuckie kills the family in a heinously obscene fashion.

Is Chuckie morally responsible for killing the family? It certainly seems that he is. After all, Chuckie is as psychologically continuous with Chuck as any of us is with our own past selves. Chuckie hasn’t been covertly controlled, and has a system of values which he accepts as his own. It seems that Chuckie is just as responsible for the murder as Chuck would have been had the cerebrum transfer not taken place. But this is a problem for the historicist who embraces the biological account of personal identity, because according to that account Chuck and Chuckie are different persons, since they are different human animals. Hence Chuckie – just like post-Chuck in the previous example – lacks the kind of history that historicists take to be a requirement on synchronic ownership. Thus the combination of historicism and the biological account of personal identity renders Chuckie not morally responsible for the murder.

One possible response to this objection is to try to disassociate moral responsibility from personal identity, conceived as sameness of human animal. This is the position that Olson takes:

> Purely biological continuity … is not related to moral responsibility in anything more than a purely contingent way: my merely being the same animal as someone is no reason to hold me accountable for his actions, or to hold him accountable for mine.

(Olson 1999: 58)
Olson makes this move precisely to avoid the unpalatable consequences of cases analogous to Cerebrum Transfer. In particular, in a ‘cerebrum-swap’ case, where Prince’s cerebrum is placed in Cobbler’s body (call the resulting person ‘Brainy’) and Prince’s body is destroyed, the biological view entails that Prince and Brainy are different persons. Nonetheless, Olson agrees with defenders of the psychological account that Brainy is morally responsible for Prince’s past actions. Olson’s response is to reject the orthodox view that moral responsibility presupposes personal identity – in other words, the view that an individual can only be morally responsible for her own actions.

The historicist, unfortunately, cannot follow Olson’s lead here. Remember, an account of personal identity is needed to establish what counts as the individual’s history and what doesn’t, which in turn establishes whether or not the individual in question satisfies the historical conditions on moral responsibility. Chuckie – a brand-new human animal with Chuck’s cerebrum – has no history according to the biological account: since Chuckie is not Chuck, Chuck’s history is not Chuckie’s history. Chuckie fails to satisfy those historical conditions; hence he is not morally responsible for murdering the family.

I showed in §2.5 that the combination of historicism with a psychological account of personal identity is an unhappy one: it delivers the result that post-Chuck is not morally responsible for murdering his neighbour. Both structuralists and historicists have equally large bullets to bite: the structuralist must bite the bullet with respect to Beth, whilst the historicists must bite the bullet with respect to post-Chuck. In this section, I have shown that the same is true if we combine historicism with a biological account of personal identity. That combination delivers the result that Chuckie is not morally responsible for murdering the family – again, a similarly-sized bullet to the one the structuralist has to bite when it comes to Beth. In other words, it might be intuitively plausible that Beth lacks moral responsibility, but it is also intuitively plausible that post-Chuck is morally responsible, and that Chuckie is morally responsible. The structuralist has to bite the bullet in the first case; the historicist has to bite the
bullet in one or other of the other two cases. But if historicists have to bite the bullet somewhere, then it seems that they are in no better a position than the structuralist; hence historicism is without decisive motivation, and compatibilists lack a reason to abandon structuralism in favour of historicism.

2.7 The collapse of historicism

While brainwashing cases are a problem for structuralism, historicists face similar counter-examples. These cases feature a causally deterministic individual who is ‘designed’ to A whilst satisfying all plausible current compatibilist (that is, both historical and structural) conditions on moral responsibility (that is, synchronic ownership). Source incompatibilists claim that such individuals are not morally responsible; after all, if determinism is true and there is only one possible future, a designer (with knowledge of all facts about the past and the laws of nature) can effectively ‘script’ every one of an individual’s actions by designing his zygote as she wishes. Pereboom – a source incompatibilist – claims that the following principle underlies such worries:

(O) If an agent is morally responsible for her deciding to perform an action, then the production of this decision must be something over which the agent has control, and an agent is not morally responsible for the decision if it is produced by a source over which she has no control. (2001: 47)

The designed individual lacks control over the source of her actions because she has been causally determined by events external to her, including events prior to her existence, and so is not morally responsible for her actions. Source incompatibilists conclude that historicism, and compatibilism in general, ought to be rejected because causal determinism precludes the kind of control they claim is intuitively necessary for moral responsibility. So, just as certain cases appear to motivate historicism in virtue of being counter-examples to the structuralist conditions on
moral responsibility, other cases appear to motivate source incompatibilism in virtue of being counter-examples to *all* current compatibilist conditions on moral responsibility.

Historicists have not been moved by such apparent counter-examples. Fischer (2011) argues that if an individual satisfies all the compatibilist conditions on moral responsibility (both historical and structural), then that individual *is* morally responsible. In this section, I want to examine the historicist’s bullet-biting stance. I will argue that once we examine the historicist’s implicit *methodological* commitments it is evident that her position is unstable because it entails two contradictory methodological principles. Attempts to stabilise historicism will then be shown to be unsuccessful, thereby leading to the collapse of historicism into source incompatibilism.

Recall how historicists motivate their position: they appeal to alleged control counter-examples to the structuralist conditions on moral responsibility, such as Brainwashed Beth. Despite satisfying the structuralist conditions on moral responsibility, Beth does not seem morally responsible for killing George because her newly-acquired system of values, which led to the formation of her first-order desire to kill George and the corresponding second-order volition, bypassed the system of values she developed herself. Given that it is plausible that Beth satisfies the other structuralist conditions on moral responsibility, it seems that Brainwashed Beth is a counter-example to all such structuralist conditions on moral responsibility.

Historicists enter the picture at this point. They claim that covertly controlled individuals like Beth are not morally responsible because they have not satisfied a historical condition on moral responsibility. Such historical conditions promise compatibilists a way to avoid the troublesome counter-examples that plague structuralism. Thus, there initially appears to be strong motivation for compatibilists to ‘go historical’ (Mele 2006: 176). Historicists, as noted, claim that structural conditions must be supplemented with a historical condition of some sort and, as I argued in §2.5, the only plausible historical condition is a positive one.
The sense of hope that historicism elicits in compatibilists is – or, I shall argue, should be – short-lived. Given that historicism is motivated solely by appealing to control cases, as I’ve argued, it entails a commitment to the following methodological principle:

(1) Control cases are an adequate test of the conditions on moral responsibility.

According to (1), if the intuitive verdict of a control case is that an individual is not morally responsible, then that individual is indeed not morally responsible. And if the individual satisfies a particular set of conditions on being morally responsible (i.e. conditions on synchronic ownership), then, following (1), those conditions are insufficient for moral responsibility.

There are, however, cases that appear to be counter-examples to all plausible current compatibilist conditions on moral responsibility, including the proposed historical conditions on moral responsibility. Consider the following case:

[Designed Ernie:] Diana [a powerful deity] creates a zygote Z in Mary. She combines Z’s atoms as she does because she wants a certain event E to occur thirty years later. From her knowledge of the state of the universe just prior to her creating Z and the laws of nature of her deterministic universe, she deduces that a zygote with precisely Z’s constitution located in Mary will develop into an ideally self-controlled agent [called Ernie] who, in thirty years, will judge, on the basis of rational deliberation, that it is best to A and will A on the basis of that judgment, thereby bringing about E. (Mele 2006: 188)

Is Ernie morally responsible for A-ing? It is plausible that Ernie satisfies both the historical and structural compatibilist conditions on moral responsibility. Despite this, incompatibilists claim that he is not morally responsible because his whole life was simply living out a plan set into motion by Diana. (Indeed, they claim that Ernie is not morally responsible for any of his actions.)
If we judge that Ernie is not morally responsible, this is a big problem for compatibilism. However historicists, as noted, don’t seem too worried about such cases. Fischer, for example, asks:

How could it increase the cost of compatibilism to show that a compatibilist must accept that an agent is morally responsible in a scenario that is no different than an ordinary situation in which there is no special reason to call into question the agent’s moral responsibility? (2011: 271)

Fischer is alluding to the fact that incompatibilists often use Designed Ernie as part of an argument against compatibilism – namely, the ‘Zygote Argument’, which is an instance of a manipulation argument. The second premise of this argument is that there are no relevant differences between Ernie and any causally determined individual. If we hold that Ernie is not morally responsible, incompatibilists press us to conclude that all causally determined individuals are just as non-responsible as Ernie. Fischer’s move is to reverse the order of the argument: if we hold that causally determined individuals can be morally responsible, it follows that Ernie can be morally responsible too. This is supposed to undercut the reliability of the judgement that Ernie is not morally responsible.

However, my interest is not in the success of Fischer’s reply to the Zygote Argument as such, but rather in the implications that his reply has for historicism. By arguing that Designed Ernie leads to an unreliable judgment, Fischer commits historicism to the following methodological principle:

(2) Design cases are an inadequate test of the conditions on moral responsibility.

---

58 I discuss this style of reply in Chapters 4, 6, and 7.
So, even if the intuitive verdict of a design case is that an individual is not morally responsible, this does not mean the individual is indeed not morally responsible; in other words, according to (2), our intuitive judgements about design cases do not tell us anything about the conditions on moral responsibility.

By denying that Designed Ernie leads to a costly implication for historicism, Fischer is effectively saying that Designed Ernie isn’t an adequate test. But if Designed Ernie isn’t an adequate test, why should we accept that Brainwashed Beth is an adequate test? This is a question historicists must answer but, I will now argue, cannot.

It might seem that (1) and (2) are consistent, because it might seem that design is not a form of covert control. Perhaps it does not seem that Diana controls Ernie because Diana’s activities are temporally distant from Ernie’s actions. However, the same is true of Beth: her manipulators do not control her from moment to moment, but they still control her in an important sense. To be sure, Beth’s manipulators are temporally closer to Beth than Diana is to Ernie. But both Beth and Ernie are in causally deterministic worlds with only one possible future. The temporal distance between controller and controllee is therefore irrelevant. So, while Beth is controlled via brainwashing in conjunction with her manipulators’ knowledge of her subsequent environment, Ernie is controlled via design in conjunction with Diana’s knowledge of all facts about the past and the laws of nature.

Put it this way. If Beth is a puppet, then so is Ernie. Whereas Beth’s strings only stretch to the day before her action, Ernie’s strings stretch all the way to events prior to his existence. But of course a puppet master still has control of her puppet no matter what length the strings connecting them are. The fact that Beth is a puppet is one plausible source of the judgement that

59 While it is not explicitly part of the case that Beth’s manipulators have knowledge of Beth’s future environment, it is certainly implied. Since it is implied, it is a relevant consideration. Another concern might be that on this way of interpreting the case, we couldn’t replace the manipulators with a brain tumour because brain tumours cannot have knowledge. This is true. But we could replace the manipulators with some other intentionless force which has local control of her environment. As I argue in Chapters 5-7, such additional factors can be combined with weak forms of manipulation, such as brainwashing and design, to constitute a form of covert control. My responses to brainwashing cases in §4.8 also explains why a brain tumour case is not a counter-example to the structuralist conditions on moral responsibility.
she is not morally responsible. And while Ernie is also a puppet, historicists are forced to accept that he is morally responsible for his actions. Historicists are therefore actually committed to the two following clearly contradictory methodological principles:

1. Control cases are an *adequate* test of the conditions on moral responsibility.
2. Control cases are an *inadequate* test of the conditions on moral responsibility.

Commitment to these contradictory methodological principles renders historicism unstable; thus historicists must reject one of them. Historicists cannot reject (1), because then they would lack motivation for their position; so they must reject (2*). But given (1), Designed Ernie shows that historical conditions are insufficient for moral responsibility, and therefore historicism collapses into source incompatibilism. What historicists need, then, is some principled way to differentiate between Cases 1 and 2 – that is, they need some way to argue that only Brainwashed Beth is an adequate test of the conditions on moral responsibility. If they could do this, they could endorse (2) instead of (2*).

Historicists could try to ground this difference on the basis of our apparently differing intuitive judgements about these cases. They could claim that while it is clear that Beth is not morally responsible, it is not clear that Ernie is not. But if this true, it really just exposes a shortcoming of Designed Ernie, namely that it is remarkably under-described. Notice that (i) we’re not told that Ernie has done something morally reprehensible, and (ii) there is no suggestion that Ernie has been changed in some way. Of course, a stipulation of Designed Ernie is that Ernie has been created by Diana, whereas Beth was modified by her manipulators. But we can easily change this. Let’s stipulate that Diana modified a pre-existing zygote. Now let’s stipulate that Diana designs Ernie to murder his neighbour, Francis. And let’s stipulate that if Diana had not intervened, Ernie would have developed into an exceptionally nice person who wouldn’t even contemplate murder. Designed Ernie is now much more like Brainwashed Beth. Let’s call
the modified Designed Ernie case ‘Designed Ernie*’. Is Ernie* morally responsible for murdering Francis? If we agree that Beth is not morally responsible, then it seems we must agree that Ernie* is not morally responsible. After all, both have been caused to do something morally reprehensible as a result of being under the control of another agent, and both would not have acted this way if it were not for factors beyond their control. Hence, it seems frankly implausible for a historicist to insist that we have different (in kind or strength) judgements about these two cases.

Potentially, though, historicists could propose a further principle that could differentiate between manipulation and design cases. Recall that Fischer argues that our judgements about Designed Ernie are unreliable because it seems the opposing judgement can be generated by generalising our judgement from an analogous ‘normal’ case in which only causal determinism obtains to Designed Ernie. Fischer claims that this shows that ‘the mental states and intentions of the distal creators of the zygote are irrelevant to Ernie’s subsequent moral responsibility’ (2011: 269). To put it more broadly:

(3) Events prior to an individual’s existence are irrelevant to her moral responsibility.

If true, this claim would explain why historicists can consider Designed Ernie* to be an inadequate test: Diana’s actions occur prior to Ernie*’s existence, and so we should not grant the judgement that Ernie* is not morally responsible argumentative weight. In other words, (3) would permit historicists to endorse (2) instead of (2*).

But (3) is false. As I noted in §2.3.3, we often look to certain events in an individual’s history to ascertain whether or not she is morally responsible for a particular action. It seems that the events prior to an individual’s existence can, at least in principle, be relevant to her moral responsibility by counting as evidence that an individual is or is not morally responsible for an action. An implication of (3) is that we should restrict our sources of evidence to events while an agent exists; but this is arbitrary. It is plausible that we could revise our attribution of moral
responsibility to an individual once we learn about events prior to that individual’s existence. For example, suppose Chad is an animal torturer who initially appears to be morally responsible for his actions. If we then learn that a scientist exposed Chad’s zygote, shortly after it was conceived, to a gene-modifying, and thereby character-changing, chemical, this could in principle constitute evidence that Chad is not morally responsible. It could, for example, serve as evidence that Chad is mentally ill at the time of action. Hence events prior to an individual’s existence are at least epistemically (and causally) relevant to an individual’s moral responsibility.

Historicists could perhaps accept this and instead claim that:

(4) Events prior to an individual’s existence are constitutively irrelevant to her moral responsibility.

The fact that events prior to an individual’s existence may be evidence that an individual does not satisfy the conditions on moral responsibility is compatible with (4). Historicists can thus potentially endorse (2) instead of (2*) on the back of (4).

But consider the following principle, which is analogous to (4):

(4*) Events prior to the time of action are constitutively irrelevant to an individual’s moral responsibility.

From this principle it follows that Beth is morally responsible and hence that Brainwashed Beth is not a counter-example to the structuralist conditions on moral responsibility. Just as the historicist can accept that design is in principle evidence that an individual is not morally responsible, the structuralist can accept that brainwashing is sometimes evidence that an agent is not morally responsible. But structuralists could claim that since it has been stipulated that Beth satisfies all the structuralist conditions on moral responsibility, she is in fact morally responsible.
for killing George. Of course, if Brainwashed Beth is not a counter-example to structuralism, then historicism is without motivation.

Historicists must therefore reject (4*). Brainwashed Beth could justify this rejection, as it allegedly shows that events prior to the time of action are constitutively relevant to moral responsibility. However, the only difference between (4*) and (4) is the temporal period which is claimed to be constitutively irrelevant to moral responsibility. But the restriction imposed by (4) – the claim that constitutive relevance only kicks in once the individual comes into existence – seems arbitrary: as causal determinism entails only one physically possible future, a controlled causally determined individual lacks the ability to undo the controller’s control because she cannot do otherwise than the controller intends her to do, and that is true no matter what the temporal distance between controllee and controller is. The fact that there is large temporal distance between Ernie* and Diana is therefore irrelevant to Ernie*’s moral responsibility. Put it this way. Just as there is no way for Beth to cut the strings between her and her brainwashers, there is no way for Ernie* to cut the strings between him and Diana. Thus, it would be arbitrary for historicists to accept (4) and reject (4*).

Historicists therefore cannot endorse (4) without thereby undercutting the motivation for their own position. Hence, historicists have no grounds for endorsing (2) instead of (2*); so they must reject either (1) or (2*). Given that historicists cannot reject (1), because otherwise historicism would lack motivation, they must reject (2*). But endorsing (1) forces them to accept the outcome of Designed Ernie* – namely, that all plausible current compatibilist conditions are insufficient for moral responsibility. Hence historicism collapses into source incompatibilism. Arguably, only source incompatibilists can maintain that control cases are – across the board – an adequate test of the conditions on moral responsibility without their position succumbing to
control counter-examples. This is because it is simply built into source incompatibilism that a morally responsible individual cannot be deterministically controlled by other agents.60,61

2.8 Conclusion

In this chapter, I have argued that historicism both lacks decisive motivation and is untenable. It therefore follows that compatibilists have to be structuralists. This, of course, leaves compatibilists with a problem: control cases, such as Brainwashed Beth. Compatibilists could insist – Frankfurt-style – that control cases are not adequate tests on the conditions on synchronous ownership. But for this move to be satisfactory, compatibilists must offer reasons why control cases are inadequate tests. The purpose of Chapters 3 and 4 (and then Chapters 6 and 7 in Part Two) is to do just that.

60 In §7.5, however, I argue that even source incompatibilism succumbs to control cases because a source incompatibilist individual can still be covertly controlled by another agent/an intentionless force.

61 For other anti-historicist arguments in this vein, see Kapitan (2000) and Watson (1999). Kapitan, though, argues that an individual's history may modify the degree of responsibility she has for an action. I think Kapitan may be making a similar error to the historicist error that I diagnose in §4.2.
Chapter 3: Authenticity without History

3.1 Introduction

While I argued in Chapter 2 that historicism is (a) unmotivated because it faces personal identity-style counter-examples which structuralism avoids and (b) untenable because it collapses into source incompatibilism, there does seem to be something alluring about historicism. That is to say, historicists do seem to capture something important about moral responsibility. I think it is because of this that historicism seems prima facie plausible, despite being an ultimately untenable position. In this chapter and the next, I argue that structuralism can incorporate the important elements of those alluring features without endorsing historical conditions, and that this helps structuralists to avoid brainwashing (and hypnosis) cases.

In §3.2, I consider a hypnosis control case which I concede shows that the standard structural conditions are insufficient for synchronic ownership. Given this, I agree with historicists that these conditions must (at least) be supplemented. I call such a supplementary condition an *authenticity* condition.\(^{62}\) While I accept that the standard structuralist conditions must be supplemented with an authenticity condition, I claim that the historicist’s error was to presume that an authenticity condition must be a historical condition. As I argued in §2.3.2, what matters is the outcome of a process, not the process itself. Hence I propose a *structural* authenticity condition – namely that a morally responsible agent must possess a particular dispositional property, which stems from having a kind of psychological coherence or harmony at the time of action, to provide a narrative explanation of an action she is morally responsible for.\(^{63}\) This authenticity condition forms part of the synchronic ownership condition of my structural-narrative view.

---

\(^{62}\) I follow Haji (1998, 2009) in using the term ‘authenticity’. My usage is meant to be similar to Fischer and Ravizza’s (1998) talk of ‘ownership’, which is how they describe the kind of condition taking responsibility is. While it might be more accurate to think the extra condition on synchronic ownership which I posit as being an ownership condition, in the sense of actions *really* belonging to an individual, I use the term authenticity to avoid confusion with my own use of the term ‘ownership’.

\(^{63}\) There are links between this psychological coherence or harmony and Frankfurt’s condition of ‘wholeheartedness’ (1988) and ‘satisfaction’ (1992). However, I lack the space to discuss this link in what follows.
In §3.3, I respond to three objections about my appeal to narrativity – in particular, I argue that being able to provide a narrative explanation does not require that an individual has a genuine history. In §3.4, I concede that it is possible, though I argue it is unlikely, that Beth satisfies my proposed authenticity condition; thus it seems that Brainwashed Beth is a counter-example to my structural-narrative view. It will be the task of Chapter 4 to show Brainwashed Beth is not a counter-example to my structural-narrative view even if Beth does satisfy its synchronic ownership conditions.

3.2 Authenticity, history, and narrativity

I think that some of us are morally responsible agents. But we didn’t become morally responsible agents overnight; rather, we went through some process which turned us into morally responsible agents from the mere individuals we once were. The fact we all go through such a process, as I said in §2.3.2, might be considered a motivation for historicism. But, as I claimed, compatibilists should consider synchronic ownership as being like cake: while a process is normally required to create a cake, it is conceptually possible that a cake could be created instantly (by scientists or some powerful being). While something like a dollar bill or an artwork requires a historical process to be authentic, this analogy suggests that being a cake is not something that necessarily requires a historical process. However, it seems that there is some sort of authenticity condition at play here: not just any object that is instantly created will be a cake; to be a cake, an object must be structured in the right sort of way. Hence there are structural authenticity conditions. I think that there is something similar with respect to synchronic ownership.

To discover what this structural authenticity condition is, I shall evaluate the process that Fischer and Ravizza (1998: 208-210) claim turns us from mere individuals into morally responsible agents: taking responsibility. One the subconditions of taking responsibility will then provide inspiration for the structural authenticity condition that I shall sketch below.
Let’s first consider a type of control case that I think shows that the standard structuralist conditions are insufficient for synchronic ownership. These are hypnosis cases. In brainwashing cases, such as Brainwashed Beth, an individual is manipulated to have a large swathe of psychological states. In hypnosis cases an individual only has a small number of psychological states induced – usually just enough so that the individual satisfies the standard structuralist conditions on synchronic ownership. The individual then acts because of those implanted states. Here is one such case:

*Hypnotised Harry.* One night Harry goes to see a live hypnotist, who offers thrills and chills to a paying audience. During the performance, the hypnotist – The Great Hypatia – asks Harry to come on stage, and Harry dutifully agrees. Hypatia tells Harry that she will cure his fear of chickens. Instead, she implants in Harry a desire to rob his local bank at noon the next day, along with the corresponding higher-order volition. The next day, Harry robs his local bank.

It might seem intuitively plausible that Harry is not morally responsible; but why? Watson might say that Harry is not morally responsible because his actions do not express a value of his. But we might easily modify the case so that Hypatia implants an appropriate value instead of (or in addition to) a higher-order volition. Such a change doesn’t seem to affect whether or not Harry is morally responsible. Fischer and Ravizza might say that Harry is not morally responsible because his action stems from a mechanism which is not moderately reasons-responsive – for instance, in counterfactual scenarios where his reasons for action differ, Harry always robs the bank. But, again, we can easily modify the case. We might stipulate that Hypatia is careful to implant a mechanism that is moderately reasons-responsive – e.g. it will not move Harry to

---

64 Such cases are found in D. Locke (1975: 104-106) and Slote (1980: 149).
action if, say, there is a chance he will get shot. Again, though, Harry does not seem morally responsible just because he acts from a moderately reasons-responsive mechanism.

Historicists, of course, will claim that Harry does not satisfy some historical condition on moral responsibility. Fischer and Ravizza claim that relevant historical condition is taking responsibility, which has the following three subconditions:

First, an individual must see himself as the source of his behaviour in the sense we have specified. That is, the individual must see himself as an agent; he must see that his choices and actions are efficacious in the world. The agent thus sees that his motivational states are the causal source – in certain characteristic ways – of upshots in the world.

Second, the individual must accept that he is a fair target of the reactive attitudes as a result of how he exercises his agency in certain contexts. Note that the individual need not have any sort of theory that explains which contexts are the ones in which he is fairly praised or blamed; he simply needs to believe that he is an apt candidate for the reactive attitudes in certain contexts, and that it is not an arbitrary matter what those contexts are. …

The third condition on taking responsibility requires that the individual’s view of himself specified in the first two conditions be based, in an appropriate way, on the evidence. (Fischer and Ravizza 1998: 210-213)

At first glance, it’s not easy to spot what’s historical about any of these subconditions. The first condition is a self-conception condition: an individual who satisfies this condition understands herself and her role in the world in a particular way. The second condition is a community condition: an individual who satisfies this condition accepts that she can be treated in particular ways depending on how she acts. Both these conditions involve individuals having particular
beliefs, but having a belief doesn’t require a particular history because manipulators could conceivably implant those beliefs. It seems, then, that whatever is historical about taking responsibility is to be found in their third condition.

According to Fischer and Ravizza, the third 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 appropriate way. But the relevant notion of appropriateness must remain unanalyzed’ (1998: 236). Since the notion of appropriateness is left unanalysed, it is difficult to assess the plausibility of this historical condition.\(^{65}\) Regardless, since I’ve argued that compatibilists cannot endorse historical conditions, we can set aside this condition. As I want to focus on Fischer and Ravizza’s first condition, I will also set aside their second condition.\(^{66}\)

According to Fischer and Ravizza (1998: 208-211), the self-conception condition is satisfied through a successful ‘moral education’: we encourage children to come to see themselves as the sources of their actions, and eventually they see themselves this way – that is, they come to see themselves as agents. This condition will not help avoid hypnosis cases. We could stipulate that Harry satisfies the condition because it only seems to require that Harry have a particular belief, and Hypatia could easily implant that too. So Fischer and Ravizza’s self-conception condition doesn’t seem to help. But it is a step in the right direction. The mistake is in positing a self-conception condition that can seemingly be satisfied by implanting a few beliefs. I think that an individual must have a certain sort of conception of herself to synchronically own her actions. But the self-conception condition I shall claim is necessary for

---

\(^{65}\) One noticeable thing is that it appears to be a negative historical condition, along the lines of Mele’s (2006) or Haji’s (2009) respective ‘no bypassing’ conditions. This condition, after all, tells us what histories cause the first two conditions to not be satisfied, and that is a negative condition. Unlike Mele and Haji, though, Fischer and Ravizza seem to require a morally responsible agent has a past. Thus, it seems that Fischer’s brand of historicism, while being negative in some sense, avoids the charge that it is unmotivated, which I pressed on Mele’s and Haji’s views in §2.5.

\(^{66}\) Fischer (2011: 187-192) in response to a criticism of Mele’s (2000) concedes that this condition can also be abandoned.
synchronic ownership is stronger than the self-conception condition that Fischer and Ravizza propose because an individual cannot be caused to satisfy it simply by implanting a few beliefs.

Notice Fischer and Ravizza’s condition is ambiguous with respect to what sense of ‘source’ is at issue. An individual suffering from a mental illness might experience her actions as someone else’s. For example, an individual might pick up a pen but feel as if someone else has made her pick up that pen or someone else is moving her body for her. In one sense, the individual might see herself as the causal source of her action – that is, she knows that the action stems from her body or occurs in her mind (for mental actions). In another sense, though, the individual does not see herself as the source of her actions because such actions do not feel like hers; rather, they feel like someone else’s. We have such a feeling of ownership for most of our actions. They feel like actions of ours rather than things that happen to us.

I contend this difference isn’t just a feeling. The feeling occurs, I claim, because we are able to offer a distinctive sort of explanation for sane actions that we cannot provide for insane actions. Suppose John steals someone’s hat during a psychotic episode. John will be unable to explain why he performed this (insane) action in the same way I could explain a sane action of mine. It seems the best John could offer is something is something trivial like: ‘I wanted to do that at the time’. For sane actions, however, we are usually able to offer a deeper explanation, one which puts that action into context with our character, our past actions, and/or our planned future actions.

This sort of explanation is a narrative explanation. According to Schroer and Schroer (forthcoming: 11), a narrative explanation is an explanation of an action or a mental state which takes the form of a story. These stories do not themselves need to match the quality or coherence of a professionally-written story; it is only required that these individual stories or narrative explanations conform to the ‘approximate’ logic of a story, a story that involves the

67 Maybe John could offer some ‘story’. But, as I’ll argue later, he won’t be able to offer a sufficiently intelligible story to count as a narrative explanation.
subject reflecting upon her character, how that character has been shaped by past events, and
how it will continued to be shaped by future events. It does not need to match, in quality and
coherence, a story written by a professional author’ (Schroer and Schroer, forthcoming: 13).
What is distinctive about a narrative explanation is that it features some self-reflection and it
‘connects that mental state/action to the subject’s sense of himself, of where he has been, and where he is
going (or trying to go)’ (Schroer and Schroer, forthcoming: 13) For example, suppose I gave some
money to a homeless person. I might provide the following story:

I gave money to the homeless person because I’m the sort of person who feels bad
for the needy. I feel bad every time I walk past a homeless person and don’t give
money to them (which is often), and, after having just bought myself something to
eat, I happened to have some spare change in my pocket when I saw this homeless
person. They also had a dog. I like dogs. I like dogs because I had dogs whilst
growing up, and this might be why I’m inclined to help those with dogs. So I gave
the homeless person some money.

The story-like form of this explanation makes it a narrative explanation. Only an individual with
a particular sort of self-conception could provide such an explanation of her actions. I claim
being able to provide such explanations is necessary for synchronic ownership.

Given that Harry has had the desire/value/mechanism to rob the bank implanted, it
seems that he must not be the sort of person who robs banks. If he was the sort of person who
robbed banks, he would not need to have such desires/values/mechanisms implanted because
they would already be there. Thus, since Harry is not the sort of person who robs banks, he is
not going to be able to provide a narrative explanation of his action of robbing the bank.

To see this, put yourself in Harry’s shoes. Imagine that you rob a bank one day.
Presumably, you are not the sort of person who normally robs banks (if you are, then imagine
you have just performed an action you think you would never perform). How would you explain the fact you did this to yourself and others? While you will likely have some memory of your deliberation process in the run up to robbing the bank, this memory is going to seem very alien to you. The only explanation that I think you could come with is something trivial like: ‘I wanted to rob the bank’, and even that seems a bit of a push. Furthermore this doesn’t really offer a story about why you robbed the bank; it only tells us that you acted because of a particular desire.

What these explanations are missing is self-reflection that puts the action in context with respect to our characters, past actions, and/or planned future actions. The explanation offered above only explains why you qua individual robbed the bank – namely because you had a particular psychological state or states. On the other hand, an explanation which contained self-reflection – what I’ve called a ‘narrative explanation’ – would explain why you robbed the bank (i.e. the fact it stems from a psychological state) and how that makes sense with respect to your character, past actions, and/or your plans for the future. Such an explanation would thereby make sense of how that action is an action of yours, in a deeper sense that I have been trying to capture.

The reason you will not be able to provide a narrative explanation for robbing the bank is because you are not the sort of person who robs banks, and so robbing this bank makes no sense with respect to your character, past actions, and/or planned future actions. More precisely, you lack all the relevant psychological states (including beliefs and memories) to be able to provide a narrative explanation of this action. The fact that you robbed the bank is going to seem like a blip in your otherwise good record. Given this, it seems that you would likely attribute your action of robbing the bank to some sort of temporary psychosis, other psychological illness, or even unexplained intoxication. This suggests that your psychological life at the time of action was abnormal in the sense that you could not understand your thought processes in the run-up to robbing the bank, and we normally understand why we think the things we do even when we
think very strange things. Likewise, if we assume that Harry is not the sort of person who robs banks (like us), it seems that he will not be able to offer a narrative explanation because he lacks the requisite beliefs, desires, values, and memories to be able to offer such an explanation. This is because while Harry has some beliefs, desires, and values conducive to him robbing the bank (implanted by Hypatia), the majority of his beliefs, desires, and values are not conducive to him robbing the bank. His implanted psychological states will therefore conflict, in some sense, with his pre-implantation psychological states.

Narrative explanation should therefore play a role in an authenticity condition. I claim that being able to provide narrative explanations of our actions is necessary for synchronic ownership. So, according to my structural-narrative view, it is not simply enough to satisfy the standard structural conditions. I accept that these conditions are, as Hypnotised Harry shows, insufficient for synchronic ownership. But I think that they come close to making an action a free action. We might, then, call actions that are performed by individuals who satisfy the standard structural conditions (or one of them) free* actions. An authentic free* action is a free action. I propose the following authenticity condition:

A free* action $A$ is authentic if and only if the individual who performs $A$ is able to provide a narrative explanation of $A$.

Of course, a critic could claim that Harry (for example) could be implanted with sufficient beliefs, desires, and values so that he could provide a narrative explanation of that action. But narrative explanations, as noted, take the form of a story. Harry would have to be able to explain why he robbed the bank and how that made sense with respect to his character, past actions, and/or future actions. It doesn’t seem like a few beliefs, desires, and values would be sufficient

\[\text{And even in cases where we might not understand why we are thinking the things we are, these do not normally move us to action. If they did, I think that we would not be morally responsible – in the basic desert sense – for those actions.}\]
to accomplish that; rather, to be sufficiently intelligible to count as a narrative explanation, Hypatia would have to implant Harry with *a lot* of beliefs, desires, and values (more on this shortly).\(^6^9\)

Note that *providing* a narrative explanation is not relevant to synchronic ownership; rather, it is being *able* to provide a narrative explanation that matters. As Schroer and Schroer write:

> You can think of the “possession” of a narrative explanation for some mental state/action as being underpinned by the subject’s possession of (or the instantiation within the subject of) a complex dispositional property. Understood along these lines, the question of whether a person stage could possess a narrative explanation for some of its mental states/actions translates into the question of whether a person stage could instantiate the appropriate complex dispositional property, a property that, if triggered, would result in the subject offering a story-like explanation of some of those mental states or actions. (Schroer and Schroer, forthcoming: 11)

Thus, it doesn’t matter whether an individual actually provides a narrative explanation for some action; rather it matters whether an individual *can* provide such an explanation. In other words, what matters to synchronic ownership is that an individual forms a particular disposition when she acts. And dispositions (or dispositional properties) cannot simply be implanted or induced piecemeal. To compare, salt has the dispositional property of *being soluble* because it is disposed to dissolve *when* it is in water. But it isn’t conceivable that another object or substance have this disposition without also sharing salt’s chemical *structure*. So one can’t make a substance *soluble* without making very major changes to its underlying chemical structure.

---

\(^6^9\) Schechtman (1996: 114-119) proposes an articulation or intelligibility constraint on narratives, which also applies to narrative explanations. While I think Schechtman provides a good (though far from perfect) account of intelligibility, I think that we have an intuitive understanding what an intelligible narrative explanation amounts to that is sufficient for my current purposes.
Analogously, I claim that underlying the possession of a disposition to provide a narrative explanation is a certain sort of psychological structure, and that the correct sort of psychological structure is a coherent or a harmonious one. That is to say, an individual forms the disposition to provide a narrative explanation for an action $A$ if she has a coherent or harmonious psychology when she $A$'s. What exactly is it to have a coherent or harmonious psychological structure? It is to have whatever psychological structure is required for one to possess the relevant disposition. Again, the analogy here is with solubility: we know that solubility requires having a certain kind of chemical structure, but we may not be able to specify independently what exactly that structure is: it’s just whatever structure underpins the disposition to dissolve in water in certain circumstances.

Of course, we might be able imagine that Hypatia implants Harry with so many psychological states such that he is able to provide a narrative explanation of the action he is hypnotised to perform (i.e. robbing the bank). But I contend this would turn this hypnosis case into a brainwashing case, because it requires that Hypatia implant Harry with so many psychological states. I set this worry aside until §3.4.

We can test this account of authenticity in the following way. Let’s suppose that Harry is the sort of person who robs banks. That is, he robs banks regularly and he plans to rob banks in the future. We might also suppose that Harry would not have robbed the bank on this occasion if Hypatia had not hypnotised him to have the effective desire/motivation to do so.\footnote{Because Harry is the sort of person who robs banks, it doesn’t make much sense to say that Hypatia implants the desires/values/mechanisms to rob the bank because, given that he’s the sort of person who robs banks, those things must already be present in some form. The sort of case I have in mind here is like Shabo’s (2010) Ego Button case, which I discuss in §7.2. In this case, the manipulators “ramp up” an individual’s (already present) egoistic reasoning processes. We might imagine something analogous occurring in this case too.} Given that Harry is this sort of person who robs banks, it seems plausible that he could provide a narrative explanation for his action robbing of the bank, even though Hypatia implanted the relevant desire/motivation. So it might seem that my proposed conditions are in fact insufficient for synchronic ownership.
It seems to me, however, that in this scenario Harry is morally responsible. He is the sort of person who robs banks on a regular basis, when he robs the bank he satisfies the standard structuralist conditions, and it is in fact incidental how he came to acquire the specific desire to rob this bank. Harry could have, for example, acquired this desire by reflecting upon a motivational poster. In such a story, I find that Harry is morally responsible; so I am inclined to think Harry is morally responsible even if another agent is responsible for Harry acquiring the proximate desire (or motivation) to rob the bank. Of course, some might worry that because Hypatia implants a desire into Harry, it is Hypatia, and not Harry, who is morally responsible for the bank being robbed. But this assumes that two individuals cannot be fully morally responsible for the same action/event. Compatibilists (as I argue in §7.4) should reject this view. Moreover, anyone who accepts the possibility of joint action should reject this view. Hence, even if we think Hypatia is morally responsible, this doesn’t exclude Harry being morally responsible too.

Narrativity has done a lot of heavy lifting in this section, but it is also controversial. In the next section, I outline and respond to three potential objections to my structural-narrative view.

3.3 Three problems with narrativity

While some philosophers find narrativity very useful, talk of narrativity makes some philosophers suspicious. In this section, I’ll first outline three problems and respond to them.

First, some might think that to be able to provide a narrative explanation an individual must have a history. In other words, narratives are essentially historical. This would be problematic for my view. After all, I’ve claimed that being able to provide narrative explanations is necessary for an individual to synchronically own her actions.

This is a worrisome objection. However, while narrative views of persons require a sense of ‘narrative’ that is essentially historical, there is nothing about narrative explanations which
requires that individuals have a past. It is conceptually possible for an individual to be created instantaneously or artificially with the psychological life of a fully developed human being. Even for her first action, such an individual can plausibly provide a narrative explanation of it. She will be able to do this if she is created with a full psychology’s worth of beliefs, desires, values, and (false) memories, and that psychology (or her psychological structure) is coherent or harmonious. The fact that some or all of her memories might be false doesn’t matter: her narrative explanations will still explain her actions with respect to character, (false) past actions, and planned future actions.

Reflect again on the cake analogy. Normally, a history is required to make a cake, but that process can be circumvented in principle. The same is true with morally responsible agents: normally it takes years to go from a mere individual to a morally responsible agent, but that process can also be circumvented in principle. So, the fact that being morally responsible for an action requires being able to give a narrative explanation of that action does not require that a morally responsible agent has a real history that her narrative explanation will touch upon (though morally response agents will normally have real histories). All that is required is that an individual conceives of herself as having a history. And that is secured if an individual has a coherent or harmonious psychology. It doesn’t matter how she came to acquire this coherent

---

71 Narrative views of persons include Schechtman (1996, 2011) and Schroer and Schroer (forthcoming). See Strawson (2004) for a critique of narrative views, and for replies to Strawson see Schechtman (2007) and Schroer and Schroer (forthcoming). While I don’t discuss Strawson’s critique here, my minimal appeal to narrativity avoids the worries he has with these views.

72 Some might worry, following Davidson (1987), that instant or artificial individuals, since they lack histories, will have no thoughts, and therefore are not morally responsible because they are not even agents. Zimmerman (1999), however, has argued that such worries, which stem from an adherence to externalism about mental content, can be circumvented without rejecting externalism about mental content. We need only suppose that the instant or artificial individual has been created by an individual with a past. That individual can therefore ‘pass on’ (so to speak) the content of their thoughts. Cf. also McKenna’s (2004, 2012a, 2012b) ‘Suzie Instant’ case, which features an individual brought into existence psychologically identical to a fully develop human adult.

73 Suppose the world was created five minutes ago, including us and all our memories. We would still be able to provide narrative explanations for our free actions even if this were true. The fact our memories would be false doesn’t matter because they still help us make sense of our actions. See Zimmerman (1999) for more discussion of this sceptical scenario and its relation to the historicism/structuralism debate.

74 An instant individual who was told, shortly after coming into existence, that her history wasn’t real and that was created with false memories, would still in the relevant sense conceive of herself having a history. Even though she knows her memories are false, she can still use these false memories to make sense of her present actions.
psychology; it only matters that she has such a psychology because that means, when she acts, she forms the disposition to provide a narrative explanation of her actions. In a sense, a narrative explanation is reflective of a coherent psychology, not of an individual’s actual history. Hence narrative explanations do not require that an individual has a history.

Second, we typically hold individuals morally responsible for their ‘out of character’ actions – that is, actions for which they seem unable to provide narrative explanations. For instance, Cooper might not be the sort of person who usually harms animals, but on this occasion he found himself doing so. As long as Cooper acts ‘freely’ when he acts ‘out of character’, it seems we would still consider him to be morally responsible for his actions.

We need to be careful about what we mean by ‘out of character’ actions because there are at least three senses of what could be meant. One sense suggests that there is a pathological cause. For example, an individual might act ‘out of character’ due to a psychotic episode. In another sense, a once tight-fisted friend might one day decide to be generous by, say, buying a round of drinks. In the former case it seems true that the individual’s actions are genuinely out of character because the individual cannot provide narrative explanations for those actions, and hence they do not stem from a coherent psychology. Of course, we do not consider psychotic individuals to be morally responsible for their actions; so the fact that such individuals do not come out as morally responsible on my view is a positive thing.

The second sort of case isn’t a problem because our tight-fisted friend’s actions are not, I contend, genuinely out of character. Rather, as there is no underlying pathology, it is acceptable to say that our friend’s generous action informs us that they are not as tight-fisted as we suspected; in other words, her actions inform us of her character because we are now aware of at least some occasions in which she is willing to spend her own money. Thus, it seems we call such actions ‘out of character’ in an epistemic sense – that is, out of character from what we believed their character to be, but not genuinely out of character.
The third sort of case is similar to a hypnosis case. These are the sorts of cases where we do something completely uncharacteristically and then think, ‘what was I thinking?’ The kind of confusion these sorts of actions leave us with suggest that they are not really ‘our’ actions – that is, in the synchronic ownership sense. The only difference between this sort of out of character action and a hypnosis case is that (hopefully) there are no nefarious manipulators secretly inducing us to have certain desires or values. As such, when we perform such ‘out of character’ actions we do not synchronically own those actions because they are not authentically ours. Hence, there is no problem with respect to ‘out of character’ actions for my structural-narrative view.

Third, it seems that few of us consciously think of our actions as part of a narrative; hence, it might be claimed that it cannot be a condition on synchronic ownership that we provide narrative explanations of our actions. In other words, my synchronic ownership condition should be rejected because it makes unrealistic claims about the psychological lives of persons.

The first point is to emphasise is that actually providing a narrative explanation is not required for synchronic ownership. All that is required is being disposed (or able) to, and having a coherent or harmonious psychology at the time of action secures that. My claim is that when an individual has such a psychology, she is able to provide a narrative explanation of her free actions, though she might never actually provide such an explanation.

It’s true that we often don’t consciously explain our free actions, to others or even to ourselves. But, according to Schechtman (1996: 114-115), narratives do not need to be consciously considered. As Schroer and Schroer put it:

… the test for possession [of a narrative explanation] is counterfactual: if queried, [an individual] could produce an explanation of the relevant mental states/actions that
lays bare their significance in virtue of embedding these states/actions in a story (or stories) about herself and her life. (Schroer and Schroer, forthcoming: 11)

There is therefore no worry that we do not consciously ‘write’ narrative explanations of our actions. What matters is that if asked we would provide a narrative explanation about a particular free action. Of course, we might not be able to provide such explanations while we are acting. But this is fine because I claim that it follows from having coherent or harmonious psychology at the time of action that an individual forms the disposition to provide a narrative explanation of her action. As long as the individual forms this disposition (that is, instantiates the dispositional property to provide a narrative explanation) at the time of action, there is no problem with her only being able to provide a narrative explanation after the time of action.

3.4 Brainwashed Beth

Historicists, as I discussed in Chapter 2, also use brainwashing cases, such as Brainwashed Beth, to motivate their position. In this section, I discuss whether or not brainwashed individuals satisfy my authenticity condition on synchronic ownership. I conclude that it is conceptually possible that they do, though quite unlikely, which helps to explain why they do not seem to be morally responsible.

Let’s remind ourselves of the Beth’s story. Beth is an exceptionally sweet woman, who would never even considering harming anyone. She is then brainwashed to have the system of values of a serial killer, Chuck. Upon awaking with these values, Beth forms the desire to kill her neighbour, George. Having previously been an exceptionally sweet person, Beth is surprised by her desire. But, as she judges that it is in line with her values, she does not hesitate to act upon it. Later that day, Beth kills George. Given that it seems intuitively plausible that Beth is not morally responsible and that she satisfies the standard structuralist conditions, it seems that those conditions are insufficient for synchronic ownership.
While we might judge that Beth is not morally responsible, there are some important details of this case which seem dubious. By hypothesis, Beth has the background psychological contents – that is, the beliefs, desires, memories, and all the rest – of an exceptionally sweet person, and now she has the values of a serial killer. These new values lead to the production of a desire to kill George. Mele (2013a: 169) claims that Beth would be ‘surprised by this new desire’. Presumably, she would also be surprised by her unexplained acquisition of the values of a serial killer. But merely being ‘surprised’ by her murderous desires and her serial killer values seems unlikely. It seems that any previously exceptionally sweet person would be more than just surprised by this. It seems that they would find it downright horrifying. Beth, after all, remembers being exceptionally sweet, and yet she now has the values of a serial killer and has a desire to murder her neighbour. At the very least, we would expect Beth to be severely confused by this sudden change. Indeed, it is likely that she would believe that she had a severe mental illness or she would display the symptoms of someone suffering from a severe mental illness. This seems like the most realistic interpretation of the details of this case. Beth might even attribute this nefarious change in her values to the actions of some other agent. Her attribution would be similar to how a schizophrenic suffering from thought insertion attributes a thought she is having to some other agent – that is to say, the schizophrenic thinks that someone else is thinking thoughts in her mind. The only difference being that Beth is correct: some other agent is responsible for her change of values.

Given that Beth has the values of a serial killer, she will be able to make sense of her action with respect to her character (which we can assume is constituted, at least in part, by her values). She might even be able to make sense of her action with respect to her planned future actions, if we assume that her planned future actions derive solely from the desires which arise as a result of her new system of values. But, because Beth has the beliefs and memories of an exceptionally sweet person, she is not going to be able to make sense of her action with respect to her past actions. It is because of this that I think that Beth will be convinced that she is
suffering from a mental illness. To see this, imagine that you are implanted with values which fundamentally oppose the values you have, and that your prior values are erased. As a result of having these values, you then do something to which you are (or were previously) opposed. Would you be able to provide a narrative explanation of your action? I don’t think I could, and I suspect I’m far from alone in finding that.

Moreover, imagine Beth’s attempt at a narrative explanation. She might wonder why she killed George. When she reflects, she will accept that she wanted to kill George and that killing George made sense with respect to her values (i.e. those of a serial killer). She might even be able to make sense of her action with respect to her planned future actions (if, as noted, those are constituted by the desires which arise from her new system of values). But suppose she tries to make sense of her action with respect to her past actions. Given that Beth was an exceptionally sweet person all her past actions will be those of an exceptionally sweet person, and this will be all she remembers. There will therefore be a disparity between her current conception of herself and how she came to be that way, as her memories suggest that she shouldn’t be the sort of person she now is. This disparity seems like it would render Beth’s narrative explanation unintelligible because her narrative would be muddled between memories and beliefs associated with being an exceptionally sweet person, having the values of a serial killer and having plans for future heinous actions. If Beth had undergone some sort of transformative experience, such as a religious conversion, she might have been able to bracket her past actions as those of her ‘old self’. But she hasn’t; she has no consciously available reason to explain her sudden change in values. So, it seems plausible that Beth does not satisfy my authenticity condition when she acts even if she satisfies the standard structuralist conditions; this therefore explains why Beth is not morally responsible for killing George.
Of course, in response to arguments like the one given above, Mele insists that it is conceptually possible that Beth is a coherent agent when she acts. That is, Mele claims it is conceptually possible that Beth will merely find it ‘surprising’ that she has new values and will thereby be able to make sense of why she kills her neighbour. If Mele is correct, then it seems that Brainwashed Beth is prima facie a counter-example to my structural-narrative view. After all, if Beth is a coherent agent, then she should be able to provide a narrative explanation of her action of killing George. Rather than argue against Mele’s claim about what is conceptually possible and what is not, I will assume that Beth is a coherent agent when she kills George. Hence, Brainwashing Beth is a prima facie counter-example to my structural-narrative view.

I should note that I think that one reason why Beth seems to lack responsibility is because the natural reading of Brainwashed Beth is that she is psychologically incoherent at the time of action, and I think natural readings influence our judgements about cases much more than stipulations about them do. Nevertheless, I will show that, even granting Mele his claim, Brainwashed Beth is not a counter-example to my structural-narrative view. To show this, I will have to outline the diachronic ownership condition – that is, the condition on moral responsibility over time – of my structural-narrative view. As most assume that diachronic ownership is a simple matter of personal identity, I will first have to show that personal identity is not the diachronic ownership condition. Before turning to this task, I must respond to an objection.

3.5 Objection

In §2.7, I argued that historicism is methodologically inconsistent, and thus collapses into source incompatibilism, because historicists accept that one sort of control case (i.e. brainwashing cases) is an adequate test on the conditions of synchronic ownership, whilst denying that another sort

---

75 Mele said something like this to me in response to my reply to his paper at a workshop on manipulation arguments in Budapest, 2012. Mele (2006: 178) also says something along these lines in response to Dennett’s (2003) argument that Beth’s manipulators would have to also implant a range of pseudo-memories in order for Beth to be a coherent agent.
of control case (i.e. design cases) are an adequate test on the condition of synchronic ownership. In other words, historicists claim that one sort of control case warrants positing further conditions on synchronic ownership, whilst claiming that another sort of control case doesn’t. A similar charge could perhaps be brought against my structural-narrative view. After all, I’ve argued that one sort of control case (i.e. hypnosis cases) forces us to posit further conditions on synchronic ownership, while I’ve implicitly conceded (and shall argue in Chapter 4) that another (i.e. brainwashing cases) does not. However, this is not a problem because there is a significant difference between these two sorts of control case.

Kane (1996: 64-65) distinguishes constraining control from covert nonconstraining control. In cases of the former, Kane claims that the control is constraining because victims of it know they are being controlled. For example, an individual with a gun to her head knows that she is being controlled by the person holding the gun. In cases of the latter, individuals do not realise they are being controlled because their manipulators have (unbeknownst to the victim) have caused them to have the psychology they now have; thus victims of covert nonconstraining control believe that they are acting entirely of their own free will.

I agree there is a distinction here, but I disagree with Kane’s assessment that it is the knowledge of being controlled which is the constraining factor. It seems to me that it is the effect of that knowledge, rather than the knowledge itself, that is constraining; it is constraining in the sense that it has a noticeable effect on an individual’s psychology – that is, by making her psychology ‘abnormal’ in a certain sense. This makes room for another type of control: covert constraining control. Just as a person with a gun to her would have an abnormal psychology – e.g. intense emotions and confusing thoughts – victims of covert constraining control might also have intense emotions and confusing thoughts, and, due to the covert nature of the control they are victims of, they might also not understand why they are acting as they are or they might not be able to understand their thoughts prior to them acting.
This is the sort of thing that happens to victims of hypnosis: because they have been implanted with desires/values/mechanisms that lead to actions that they cannot provide narrative explanations for, those implanted psychological states create a kind of psychological incoherence or disharmony which is noticeable to the hypnotised individual – namely, they cannot understand what they are thinking or even why they are thinking it. Thus hypnosis is a mode of covert constraining control. (If we think brainwashing renders an individual incoherent, then it is also a mode of covert constraining control. But since I’ve assumed brainwashed individuals are coherent, I will classify it as a mode of covert nonconstraining control.\(^\text{76}\) I avoid the charge of methodological inconsistency because I only posit further conditions in response to cases of covert constraining control. In response to covert nonconstraining control cases, as we’ll see in Chapters 4-7, I argue that victims of such covert control (as long as they satisfy my synchronic ownership conditions, of course) are morally responsible.

\(^{76}\) Also, my argument against historicism in §2.7 implicitly assumes that brainwashing is a mode of covert nonconstraining control. We could distinguish between two sorts of brainwashing case – i.e. one where individuals are rendered incoherent and one where they remain coherent when they act. The latter would be a nonconstraining form of brainwashing, whereas the former would be a constraining form of brainwashing. Including the distinction in what follows would needlessly overcomplicate things, however.
Chapter 4: Diachronic Ownership and Brainwashing

4.1 Introduction

In Chapter 3, I gave an initial response to brainwashing cases: I argued that brainwashed individuals likely do not satisfy my structuralist authenticity condition. However, I conceded that it is conceptually possible that a brainwashed individual satisfies this condition. In this chapter, I respond to brainwashing cases. These cases pose a problem for my structural-narrative view, and structuralism in general, because they allegedly show that my synchronic ownership conditions (one of the standard structuralist conditions and my authenticity condition), and the standard structuralist conditions, are insufficient for moral responsibility, and hence that my view, and structuralism in general, is false.\textsuperscript{77} I argue that there is an explanation why such brainwashed individuals, if they are coherent agents, are morally responsible. Hence such cases do not show that my structural-narrative view is false. Thus I show that compatibilists can resist brainwashing cases.

So far I have only considered structuralism to be an account of synchronic ownership.\textsuperscript{78} In this chapter, however, I wish to emphasise the relevance of diachronic ownership. While synchronic ownership refers to the conditions which must obtain for an individual to be morally responsible at the time of action, diachronic ownership refers to the conditions which must obtain for an individual to be morally responsible at t\textsubscript{2} for an action performed at t\textsubscript{1} – that is, moral responsibility over time. The orthodox view in the literature is that diachronic ownership is a simple matter of personal identity: an individual at t\textsubscript{2} is morally responsible for a free action performed at t\textsubscript{1} if and only if the individual at t\textsubscript{2} is the same person as the individual who

\textsuperscript{77} Another way to put, which is gaining increasing prominence in the debate, is that control counter-examples display a position’s ‘theoretical cost’. The idea is that when a view is sufficiently costly, we should reject it. I use this phrasing in Part Two.

\textsuperscript{78} Frankfurt’s and Watson’s conditions may be considered sufficient in conjunction with relevant epistemic conditions.
performed the free action at \( t_i \). How is diachronic ownership relevant? Well, as I argue in §4.3, the *locus* of moral responsibility – that is, the entity to which the property of *being morally responsible for* \( A \)-ing ‘attaches’ – is determined by whatever we take the diachronic ownership condition (DOC) to be. Since the orthodox view is that personal identity is the DOC, we normally assume that the locus of moral responsibility is the ‘whole’ (i.e. temporally extended) individual (and whatever this amounts to depends on our favoured account of personal identity). I then argue that personal identity is not the DOC. In §4.4 I argue that biological continuity (or any form of substance continuity) is not the DOC. In §4.5 I argue that neither unique psychological continuity (the criterion of the psychological approach) nor non-unique psychological continuity is the DOC. In §4.6 I argue that neither ‘mere’ nor ‘strong’ psychological connectedness is the DOC. Having rejected these various candidate DOCs, in §4.7 I present my positive proposal by outlining the DOC of the structural-narrative view – namely ‘narrative connectedness’. In §4.8 I finally answer the question at hand: I deploy the proposed DOC to argue that brainwashing cases are not counter-examples to my structural-narrative conditions on synchronic ownership. I also offer a preliminary response to design cases.

### 4.2 Historicism and diachronic ownership

Before I begin, I wish to make a further observation about historicism. Consider part of McKenna’s (2004: 183) response to Brainwashed Beth. He concedes that there is an intuitive

---


80 I should note that I only focus on so-called ‘complex’ views of personal identity – that is, accounts which posit informative and noncircular conditions for personal identity over time – and I do not discuss so-called ‘simple’ views, which deny that there are informative or noncircular conditions for personal identity (e.g. Swinburne 1971, 1984; Merricks 1998; Baker 2000). One reason I don’t discuss these views is that my arguments against complex views of personal identity being the diachronic ownership condition apply equally well to simple views. The other reason is that I find simple views to be implausible, though since this thesis isn’t about personal identity over time I won’t argue that this is the case. For a criticisms of simple views see, for example, Zimmerman (1998) and Olson (2013).
difference between Beth and Chuck (the serial killer whose values Beth has had implanted). But he argues that the difference is not that Beth is not morally responsible for murder and Chuck is; rather, it is the fact that she is *not* morally responsible for having the values she has, whereas Chuck *is* morally responsible for having the values he has because he developed himself into the wicked person he is. Hence, we can explain why there is an intuitive difference between Beth and Chuck without conceding that Beth is not morally responsible for murder.

This seems correct to me: Beth is certainly not morally responsible for becoming the sort of person who murders people, whereas as Chuck is. This highlights another historicist mistake. As I noted in §3.1, historicists seem to get something correct with the requirement of an authenticity condition (that is, an extra condition on synchronic ownership beyond the standard structuralist conditions), but they mistakenly infer that an authenticity condition must be a historical condition. I think they also get something correct in positing a diachronic condition on moral responsibility. However, their mistake is in thinking that it is a diachronic condition on synchronic ownership. Since Chuck is uncontroversially responsible for more than Beth, historicists tried to craft a synchronic ownership condition to capture this claim. However, the reason that Chuck is responsible for more is that he is morally responsible for becoming the person he is, while Beth is not. In other words, Chuck is related in the right sort of way to the previous time-slices of himself who moulded him into the serial killer he now is. Beth, on the other hand, is not related in the right sort of way to the time-slices of the other individuals who turned her into a murderer; so she is not morally responsible for becoming the murderous person she now is.

I suspect that historicists presume their intuition is about synchronic, rather than diachronic, ownership because they presume that diachronic ownership is a simple matter of personal identity, and so they conclude their intuition cannot be an intuition about the DOC. In

---

McKenna is actually talking about a different case (or pair of cases), which features Beth being implanted with Ann’s values rather than Beth being implanted with Chuck’s values. As the case is structurally similar to Chuck/Beth case, what McKenna says about Ann/Beth applies to Chuck/Beth.
what follows, I argue that personal identity is not the DOC. This paves the way for a new DOC and, as I argue in the next section, the locus of moral responsibility is determined by whatever the DOC is.

4.3 The locus of moral responsibility

Consider how Mele tries to supplement his argument for historicism. He claims that we view ourselves ‘as relatively thick, temporally extended beings’ because ‘praise, blame, reward, or punishment are deserved by agents now for past deeds’ (Mele 1995: 165). His point is that it is part of our practice to hold each other morally responsible for past actions, and this is supposed to further support his argument for historicism. But this obviously doesn’t help the historicist. No structuralist will deny that we view ourselves this way and, indeed, that all of us – bar some conceptually possible beings – are temporally extended beings. This highlights, I think, that structuralists ought to explicitly take themselves to be defending a time-slice (or person-stage) account of moral responsibility according to which the property of being morally responsible for A-ing ‘attaches’ itself in the first instance to the time-slice (or person stage) that A’s. This is not to say that all structuralists do defend such an account, but rather that structuralism is only plausible if this commitment is made. We implicitly accept that the locus of moral responsibility is the whole temporally extended individual, and it is for this reason that I think structuralism has appeared implausible to many.

Why does it seem that the whole temporally extended individual is the locus of moral responsibility? Well, I claim that it follows because we implicitly hold the orthodox view that personal identity is the DOC – that is, the condition on being morally responsible over time. A consequence of this view is that an individual continues to be morally responsible for her free actions as long as she continues to exist. For example, if the psychological approach were true, an individual would be morally responsible for her past free actions if and only if she is uniquely psychologically continuous with the individual who performed those free actions. Thus, on this
view, the locus of moral responsibility is the ‘person’ – that is, the entity that the psychological approach claims we essentially are and which therefore provides our persistence conditions. On the other hand, if the biological approach were true, an individual would be morally responsible for her past free actions if and only if she is biologically continuous with the individual who performed those free actions. Thus, on this view, the locus of moral responsibility is the ‘human animal’ – that is, the entity that the biological approach claims we essentially are and which therefore provides our persistence conditions.

This point is quickly made, but it is of the utmost importance with respect to my defence of structuralism. As we’ll see in §4.8, if the orthodox view were true structuralism would have no hope of resisting brainwashing cases. However, once it is clear that personal identity is not and cannot be the DOC, structuralism will find new hope as the replacement DOC determines that the locus of moral responsibility is not the ‘whole’ temporally extended individual, and this will provide structuralists with the resources to resist brainwashing cases.

I should emphasise that I only argue that different views of personal identity cannot be the DOC; I do not comment on the coherence of those views or their adequacy as accounts of personal identity.82 Moreover, I remain neutral on whether a particular account of personal identity adequately underlies our other practical concerns, such as survival or self-concern.83

---

82 Sider (2001) argues that there is no fact of the matter which criterion of personal identity is true. This would be a worrying result if the orthodox view is true because it would mean that there is no fact of the matter about whether an individual is morally responsible for some past free action. As I reject the orthodox view I am able to side-step Sider’s argument and the problems that follow from it. Indeed, Sider’s argument provides another reason to favour the rejection of the orthodox view.

83 Often a particular account of personal identity is motivated by how well it underlies a particular practical concern (e.g. Locke used moral responsibility in this respect), and then it is criticised because it fails to underlie another practical concern (e.g. Schechtman 2010 criticises Locke’s accounts because they fail to capture our animalistic practical concerns). Shoemaker (2007) argues for a pluralist approach to the identity relations which underlie our practical concerns. Elsewhere, Shoemaker (2012) argues that personal identity is not the diachronic ownership condition.
4.4 Diachronic ownership and the biological approach

My question in this section is not what makes an agent morally responsible when she performs her action – that is a question about synchronic ownership, which clearly requires a specific sort of psychology/psychological ability at the time of action – but rather what makes an agent morally responsible for an action at a later time; in other words, I want to ascertain what the DOC is.\textsuperscript{84}

The orthodox view, as noted in §4.3 is that diachronic ownership is a matter of personal identity. As I discussed in §2.6, a plausible substance-based account of personal identity is the biological approach or ‘animalism’.\textsuperscript{85} On this view, persons are essentially human animals. This means that although most human animals have psychologies, we do not persist through time on the basis of our psychologies. This is because we can lack the property of being a person and yet continue to exist. However, we could not lack the property of being a human animal and yet continue to exist; hence, this provides our persistence condition. So, I am the same person as some other person if and only if I am biologically continuous with that person.

An obvious complaint about this approach to personal identity is that it is not about personal identity at all, but rather about human identity.\textsuperscript{86} Olson is well aware of this sort of criticism, but it doesn’t faze him. Part of his argument for the biological approach is that the psychological approach has become convoluted in its attempt to provide to a criterion of numerical identity and to underlie our practical concerns, such as moral responsibility.\textsuperscript{87} Despite defending animalism, Olson doesn’t think we aren’t persons. On the contrary, he thinks we are (at least for most of our lives) persons but just not essentially so. As human animals, it is possible

\textsuperscript{84} Although I focus on the biological approach, the points I make in this section can be applied to any substance based account of personal identity – such as bodily (Thompson 1997), soul (Swinburne 1984), or brain based (Nagel 1986) account of personal identity. Of course, one might interpret ‘souls’ as being an essentially psychological substance (whatever that means); if so, then this kind of soul view succumbs to the problems I argue psychological continuity theories have in §4.5.

\textsuperscript{85} Other animalists include van Inwagen (1990), Cassam (1993), and Johannson (2007). As Olson’s view is most well developed, I will focus on it in the following. While there are differences between each of these views, the differences are unimportant in what follows.

\textsuperscript{86} Cf. DeGrazia (2005).

\textsuperscript{87} Schechtman (1996: 51-60) makes a similar criticism of psychological continuity theories, though she doesn’t abandon the psychological approach entirely.
for us to gain and lose the property of being a person. For instance, we don’t start our lives as persons; we start them as mere biological entities, and sometimes we don’t end them as persons because sometimes we lose our psychological capacities (for example, due to severe brain injuries). Hence, according to Olson, the property of being a person is a phase sortal – it is a property we only have for a phase of our existence. In the same way that a cat is kitten only for a phase of its life, we are persons only for a phase of ours. When a cat stops being a kitten, it only loses the property of being a kitten; it is does not go out of existence. Likewise, according to Olson, we can lose the property of being a person without going out of existence (e.g. if we suffer brain damage).

I won’t assess here whether the biological approach is an adequate account of personal/numerical identity. Rather, I want to know whether the biological continuity is the DOC. The short answer is: no. As I pointed in §2.6, this is something that Olson is well aware of. Indeed, as I pointed out above, Olson doesn’t consider himself to be in the business of explaining our practical concerns, but rather in the business of doing plain and simple nuts and bolts metaphysics; so the fact that biological continuity is not the DOC is not going to bother Olson. Despite this, it will be worth seeing exactly why biological continuity (or, indeed, any substance-based continuity) is not the DOC. Consider the following case:

Brain-wipe Barry: Barry is a healthy human adult. He has developed in a normal way and satisfies the appropriate synchronic ownership conditions on moral responsibility. On the afternoon of his twenty-fifth birthday he robs a fishmonger at gunpoint for his prize crab. After devouring the stolen crab for his dinner, Barry undergoes a ‘religious experience’ – that is, he feels as if he is being taken over by a mysterious force that he attributes to God. After his experience, Barry remembers

---

88 For criticisms in this vein see, for example, Belshaw (2010), Johannson (2007), or Zimmerman (2008).
nothing about his past life. In fact, all of his psychological states have been ‘erased’. Barry’s family find him lying on the floor, covered in crab meat, and they take him to the hospital. The doctors inform Barry’s family that he is psychologically on par with a baby. Due to their kind nature, Barry’s family decide to raise him as if he were a new-born. Twenty-five years later, Barry is once again a healthy adult. Thanks to the cutting-edge anti-aging creams and drugs, Barry still looks as if he were twenty-five years old. Barry’s family regularly comment that the ‘new’ Barry is a ‘different person’ to the old Barry because, among other things, the new Barry doesn’t even like crab and wouldn’t dream of robbing a fishmongers to get his hands on the day’s prize catch.

Is the new Barry morally responsible for the robbing the fishmongers? It seems not, because the new Barry is in no way psychologically connected to the old Barry. The only thing in common between the old Barry and the new Barry is that they are the same human animal. Notice that if biological, or any substance-based, continuity were the DOC, then, according to orthodox view that a criterion of personal/numerical identity is the DOC, the new Barry would be morally responsible for robbing the fishmonger. Hence no substance-based continuity is the DOC.

4.5 Diachronic ownership and the psychological approach
Just as it seems natural to say that psychology is in some sense relevant to synchronic ownership, it seems natural to say that psychology is in some sense relevant to diachronic ownership. Locke (1690) was the first to develop this natural thought when he outlined the earliest incarnation of the psychological approach.\(^{89}\) Contemporary proponents of the psychological approach (e.g. S.

\(^{89}\) However, as I mentioned in n.22, Strawson (2011) argues that Locke was in fact trying to provide an account of moral responsibility, and not an account of personal identity. I set aside this point in what follows.
Shoemaker 1984; Parfit 1984; Noonan 2003) have all developed their views from Locke’s. Indeed, their views are often called ‘neo-Lockean’. In this section, I will argue that neither the psychological criterion of personal identity (unique psychological continuity) nor mere psychological continuity is the DOC. To understand neo-Lockeanism in general, and this distinction in particular, we must trace the development of that view from Locke’s account of personal identity. Only then will we be able to properly assess whether a psychological criterion can serve as the DOC.

Locke is usually interpreted as proposing a memory criterion of personal identity. Roughly, a person at \( t_2 \) is identical to a person at \( t_1 \) if and only if the person at \( t_2 \) has the memories of and remembers being the person at \( t_1 \). But it is controversial whether this was actually Locke’s view. As Schechtman (1996: 107) observes, Locke never actually uses the term ‘memory’ in his discussion of personal identity, yet he uses it elsewhere in his essay. Rather, Locke used the term ‘consciousness’ to denote his criterion of personal identity. Although there might be some mild historical controversy in doing so, to avoid a protracted discussion of interpretive issues, I will assume, as Strawson (2011) argues, that Locke in fact meant something closer to psychological connectedness than he did to memory connectedness by his use of ‘consciousness’. Memory connectedness holds when an agent remembers a particular experience, while psychological connectedness is broader: it includes memory connectedness, as well as between forming an intention and acting on an intention, or forming a belief and continuing to hold that belief. The assumption that Locke meant psychological connectedness, rather than memory connectedness, by his use of ‘consciousness’ shouldn’t be too problematic, because neo-Lockeans agree that other psychological states – such as beliefs, desires, intentions, and the like – are relevant to personal identity.

---

90 It’s worth noting that despite many claims that the psychological approach is most popular account of personal identity, there are few explicit defences of the psychological approach these days.

91 Although I only discuss Parfit’s version of the psychological continuity theory, my arguments will apply to so-called ‘hybrid’ views that claim that psychological continuity is underpinned by sameness of brain (e.g. S. Shoemaker 1984) or first-personal constitutional accounts, such as Johnston (1987) or Baker (2000).
Despite the obvious influence of Locke’s view, there are only a few similarities between Locke’s view and those defended by neo-Lockeans (and that’s even assuming that Locke meant psychological connectedness by ‘consciousness’, and not memory connectedness). One similarity between Locke and the neo-Lockeans, though, is that they both believe an adequate account of personal identity should underlie our practical concerns – and, in particular, that an adequate account of personal identity should provide the conditions for diachronic ownership. Another similarity is that they hold that persons are distinct from biological organisms, because a person is essentially ‘a thinking intelligent being, that has reason and reflection, and considers itself as itself the same thinking thing, in different times and places’ (1690/1975: 39), whereas a mere biological organism has none of those capacities essentially.

Locke’s argument for his view, which neo-Lockeans endorse, stems from various thought experiments that pump intuitions about moral responsibility. Consider the following case:

For should the soul of a prince, carrying with it the consciousness of the prince’s past life, enter and inform the body of a cobbler, as soon as deserted by his own soul, every one sees he would be the same person with the prince, accountable only for the prince’s actions. (Locke 1690/1975: 44)

Is the Prince in the Cobbler’s body (let’s call him Souly) morally responsible for the Prince’s actions? Locke thinks that he is. From this he infers that Souly is the same person as Prince. Of course, we might then think that the correct criterion of personal identity is sameness of soul. However, Locke argues against this. If John has Socrates’ soul, does that mean that John is morally responsible for Socrates’ actions? If John only shares a (presumably) featureless immaterial substance with Socrates, then it seems not. Hence, sameness of soul does not ground moral responsibility. According to Locke, what grounds moral responsibility (and hence personal
identity) is sameness of consciousness (or psychological connectedness). The reason that Souly is morally responsible is not that he possesses Prince’s soul, but rather that he ‘carries’ Prince’s consciousness.

Locke’s view came under major criticism from Bishop Butler (1736/1975) and Thomas Reid (1785/1975). Both claimed that there were fundamental problems with Locke’s account of personal identity. Butler, for instance, argued that Locke’s proposed criterion of personal identity is viciously circular. (Although Butler interpreted Locke as proposing a memory connectedness criterion, Schechtman [1990] argues that his criticisms still apply to psychological connectedness. So, even though I will talk about memory connectedness in the following, bear in mind that this criticism arguably applies to other psychological connections.) According to Butler’s interpretation of Locke, I am the same person I was a year ago because I remember the actions I performed a year ago. This seems fine. But it might also be the case that I have non-genuine memories – that is, apparent memories of actions that are not mine. We thus have to distinguish between genuine and non-genuine memories, and Butler notes that we have a way to do this: we must discover whether the actions I apparently remember performing are actions I did perform. In other words, we must discover whether I am numerically identical to the individual who performed the actions I seem to remember performing in order to ascertain whether I genuinely remember performing them. Distinguishing between genuine and non-genuine memories therefore requires numerical identity; but that’s what Locke was trying to provide a criterion for. Hence, Locke’s proposed criterion for personal identity is viciously circular.

Reid’s objection seems equally devastating. He argued that since personal identity is an identity relation, it is necessarily transitive. However, psychological (or memory) connectedness is not a transitive relation. Reid’s Brave Officer case shows this. Imagine an old man is psychologically connected to a brave officer – that is, the old man remembers being the brave officer and they share some beliefs, desires, intentions, etc. Imagine next that the brave officer is psychologically connected to a young boy – that is, the brave officer remembers being the young
boy and they share some beliefs, desires, intentions, etc. But we can also imagine that the old man is psychologically disconnected from the young boy, so the old man does not remember being the young boy and they share no beliefs, desires, intentions, etc. Clearly this has an absurd consequence: the old man (x) is the same person as the brave officer (y), the brave officer is the same person as the young boy (z), but the old man is not the same person as the young boy; in other words, x=y and y=z, but x≠z. Hence, since psychological connectedness is not a transitive relation, it cannot constitute personal identity.

Despite these objections, neo-Lockeans have felt the intuitive pull of the psychological approach and have sought to develop it to avoid these objections. First, they avoid to Reid’s transitivity objection by introducing the ancestral relation of psychological connectedness: *psychological continuity*. According to Parfit (1984: 206), S is psychologically continuous with R if and only if there are overlapping chains of *strong* psychological connectedness between S and R. Strong psychological connectedness holds when one individual possess at least half of the direct psychological connections to an earlier individual that usually hold day to day in normal persons. Thus, there is a psychological criterion that avoids Reid’s transitivity objection.

Second, neo-Lockeans introduce the concept of ‘quasi’ psychological states to avoid Butler’s circularity objection. The idea is that the concept of ‘q-states’ is similar to the concept of regular psychological states, except that the possession of a q-state doesn’t presuppose personal identity. Let’s see how this works with memories. The concept of memory presupposes that a person can only remember *her own* experiences. What differentiates a memory from a delusion is that when I remember an event, as opposed to deludedly *seeming* to remember it, it is an event I actually experienced. The concept of q-memory, by contrast, does not presuppose personal identity.

---

92 I will focus particularly on Parfit’s (1984) development of neo-Lockean view, although it’s inaccurate to call Parfit a neo-Lockean because Parfit in fact argues that personal identity does not matter to our practical concerns. Developing a psychological continuity is therefore a mere foil for Parfit to develop his actual view. Despite this, Parfit has most well developed neo-Lockean account of personal identity. Other neo-Lockeans include S. Shoemaker (1959, 1984, 1999) and Noonan (2003). More recently, Parfit (2012) has argued that persons are essentially embodied minds.
identity; all that it requires is that what an agent q-remembers actually happened to someone, and that q-memory is appropriately causally connected to the original memory. For example, Audrey ate a salad one afternoon and then formed a memory of that experience. Some scientists then isolated the neural substrate of that memory and implanted it into Cooper’s brain. The idea is that Cooper will wake up with a q-memory of someone eating a salad. When Cooper remembers Audrey’s experience, it will be a q-memory. Thus, q-memories do not presuppose personal identity. The same basic technique can then be applied to other psychological states as well, such as beliefs and desires. Hence, the psychological approach avoids the circularity charge.\(^{93}\)

A further worry about for the psychological approach results from cases like the following:

_Bert’s Brain._ Bert crashes his car, injuring himself and his two twin brothers. Bert’s body is damaged beyond repair, but his is brain is miraculously undamaged. The brains of Bert’s brothers, on the other hand, are damaged beyond repair while their bodies sustain minor injuries (such as cuts and bruises). The ambulance that rescues Bert and his brothers happens to be staffed by members of a radical neurosurgery research project. They decide to remove Bert’s brain, split it in two, and then to transplant those brain halves into Bert’s brothers’ vacant skulls. The neurosurgeons call the two resulting individuals ‘Bert\(_1\)’ and ‘Bert\(_2\)’. Bert\(_1\) and Bert\(_2\) are both equally psychologically continuous with Bert.\(^{94}\)

\(^{93}\) Schechtman (1990) argues q-states (and in particular q-memories) are incoherent. While her argument has some force, Northoff (2000) has recently argued that neuropsychological research supports the concept of q-states; so he concludes we must reject Schechtman’s argument on empirical grounds. Slors (2001) defends an alternative to q-states which appeals to narrativity.

\(^{94}\) This is based on Parfit’s (1984: 253-254) ‘My Division’ case. Even if this case failed for some reason (e.g., it is in fact conceptually impossible to split a brain, to transplant it into two bodies, and then for the resulting individuals to have identical psychologies), notice that my argument against the orthodox view hinges on the diachronic ownership condition _not_ being transitive. Hence these cases are not particularly important to my view.
Does Bert survive the car accident? That is, is Bert either Bert\textsubscript{1} or Bert\textsubscript{2}? Given that personal identity is an identity relation, and so by definition must hold one-one, Bert cannot be both Bert\textsubscript{1} and Bert\textsubscript{2}. Hence, it seems that Bert is neither Bert\textsubscript{1} nor Bert\textsubscript{2} because one thing cannot be identical two things. Parfit (1984: 215) takes this sort of case to show that identity doesn’t matter; what matters is rather survival, and survival consists of ‘psychological continuity and/or connectedness’.

In this sense, then, Bert survives as both Bert\textsubscript{1} and Bert\textsubscript{2}, even though he isn’t identical to either of them. In order to preserve a criterion of personal identity, then, neo-Lockeans are forced to stipulate a non-branching clause. The end result is the following sort of criterion of personal identity:

\begin{quote}
\textit{The Psychological Criterion}: (1) There is psychological continuity if and only if there are overlapping chains of strong connectedness. X today is one and the same person as Y at some past time if and only if (2) X is psychologically continuous with Y, (3) the continuity has the right kind of cause, and (4) it has not taken a ‘branching’ form. (5) Personal identity over time consists in holding of facts like (2) to (4).
\end{quote}

(Parfit 1984: 206)

Is this criterion a plausible candidate for the DOC? It seems not. We need only stipulate that Bert is morally responsible for crashing his car, and then ask whether Bert\textsubscript{1} and Bert\textsubscript{2} are morally responsible for this action. If unique psychological continuity were the DOC, neither Bert\textsubscript{1} nor Bert\textsubscript{2} would be morally responsible for Bert’s action because they would not be identical to Bert. However, it seems intuitively plausible that Bert\textsubscript{1} and Bert\textsubscript{2} are morally responsible for Bert’s

---

\textsuperscript{95} In earlier work, Parfit (1971) clearly held that psychological connectedness is what matters. But in Reasons and Persons, he seems to have changed his mind. But it’s unclear why he thinks psychological continuity matters. For criticisms of Parfit’s identity doesn’t matter view see, for example, Johnston (1987) and Schechtman (1996).

\textsuperscript{96} What the ‘right kind of cause’ is is left open. Parfit (1984: 207) identifies three candidates, though the specifics of these are unimportant for my purposes. Moreover, the non-branching clause is clearly ad hoc. Parfit’s point, it seems, is that since identity doesn’t matter, it doesn’t matter that the psychological continuity theory is problematic.
action. Additionally, if only Bert₁ had survived, he would be morally responsible for crashing the car. But it seems arbitrary that Bert₁ or Bert₂ can avoid responsibility for Bert’s actions just because there is another individual who is just as psychologically to their past self as they are. Hence, unique psychological continuity is not the DOC.

This isn’t a problem for Parfit, of course. As noted, he thinks that personal identity doesn’t matter to survival; rather what matters are psychological continuity and/or connectedness. The next best contender for the DOC, then, is non-unique psychological continuity.

Unfortunately, however, even psychological continuity is not the right sort of relation. The following case, adapted from Parfit (1984: 302-303), should make it clear that psychological continuity is not the DOC:

*Clive-1000*. Clive belongs to a special community of people who live on a hidden island in the middle of the ocean. What is special about these people is how long they live for. It is rare for normal humans to live longer than 100 years, but in Clive’s community it is normal for people to live for over 1000 years. Some attribute this to the unique conditions on island, though no one is entirely certain why this happens. Another interesting fact about this island community is that its members are continually developing their characters and forgetting their past selves. When Clive is 1000 years old (call this time-slice of Clive ‘Clive-1000’), for example, he does not remember any of his 20-year-old self’s actions (call him Clive-20) nor does he share any other characteristics with Clive-20. Thus, although Clive-1000 is psychologically and biologically continuous with Clive-20, he has no relevant direct psychological

---

97 To reiterate, I will only be arguing that psychological continuity is not the right sort of relation for *diachronic ownership*. It might be that other practical concerns are underpinned by psychological continuity. For example, whether I’m self-interested in a future person might depend on whether I’m psychologically continuous with that person. Some practical concerns might even be underpinned by biological/substance continuity. I remain neutral on this issue in what follows.
connections with Clive-20. Indeed, Clive has gone through many different characters
during his lengthy life.

Suppose that Clive-20 A-ed and in doing so satisfied all the synchronic ownership conditions. It
seems clear that Clive-20 was morally responsible for A-ing at the time at which he A-ed. But is
Clive-1000 morally responsible for A-ing? It seems not. Since Clive-1000 bears no relevant direct
psychological connection to Clive-20, the Clives are related merely by psychological continuity
(that is, each of his daily time-slices shares more than 50% of the psychological connections
from the previous day) and biological continuity (that is, each of his time-slices is part of the
same life). To say Clive-1000 is responsible for Clive-20’s action, which we would be forced to
say if psychological continuity were the DOC, seems preposterous. Cases like Clive-1000 show
that the DOC cannot be a transitive relation. Psychological continuity is a transitive relation;
thus, psychological continuity is not the DOC.

Thus, it looks like psychological connectedness is the next best contender for the DOC.
If this were correct, it would accommodate our intuitive attributions of moral responsibility in
the Bert case: Bert₁ and Bert₂ are both morally responsible for Bert’s actions because they are
both psychologically connected to Bert. It would also accommodate our hesitancy to attribute to
moral responsibility to Clive-1000 for Clive-20’s action: after all, Clive-1000 is not psychologically
connected to Clive-20.

4.6 Diachronic ownership and psychological connectedness
Although, as we have seen, psychological connectedness seems like the most plausible candidate
for the DOC, it is in fact much too weak to ground diachronic ownership. Psychological
connectedness is the holding of particular psychological connections – for example, between an
experience and a memory of that experience, between forming an intention and acting on that
intention, between forming a belief and continuing to hold that belief. Locke, in fact, was aware
of this problem (however we interpret his proposed criterion of personal identity). Consider the following passage in which he outlines the problem and proposes a solution. He asks:

whether the consciousness of past actions can be transferred from one thinking substance to another. I grant were the same consciousness the same individual action it could not: but it being a present representation of a past action, why it may not be possible, that that may be represented to the mind to have been which really never was, will remain to be shown.

And therefore how far the consciousness of past actions is annexed to any individual agent, so that another cannot possibly have it, will be hard for us to determine, till we know what kind of action it is that cannot be done without a reflex act of perception accompanying it, and how performed by thinking substances, who cannot think without being conscious of it. But that which we call the same consciousness, not being the same individual act, why one intellectual substance may not have represented to it, as done by itself, what it never did, and was perhaps done by some other agent – why, I say, such a representation may not possibly be without reality of matter of fact, as well as several representations in dreams are, which yet whilst dreaming we take for true – will be difficult to conclude from the nature of things. And that it never is so, will by us, till we have clearer views of the nature of thinking substances, be best resolved into the goodness of God; who, as far as the happiness or misery of any of his sensible creatures is concerned in it, will not, by a fatal error of theirs, transfer from one to another that consciousness which draws reward or punishment with it. (Locke 1690, Book II, XXVII: §13)

Locke’s worry is that if one individual can in principle become conscious of another individual’s action, then the former individual will, according to Locke’s view, be morally responsible for that
action. Call this the ‘problem of loose psychological connections’. This is certainly problematic. Suppose Cooper is made ‘conscious’ of an action of Audrey’s. Perhaps scientists implant the neural substrate of a memory of Audrey’s into Cooper so that Cooper q-remembers an action of Audrey’s. It would be absurd if Cooper were then morally responsible for Audrey’s action. Thus, memory connectedness is not the DOC. The same is true if Cooper were implanted with any other psychological state. Thus, psychological connectedness is also not the DOC. Locke’s solution to the problem of ‘loose’ psychological connections is to posit that God, given his moral perfection, would never allow such an unjust transfer of consciousness. But that reply is unsatisfactory. Whether or not God exists, it does not seem like a particularly convincing strategy to invoke him every time our view runs into problems.98

Thus, mere psychological connectedness is not the DOC; but what about strong psychological connectedness? This holds when one individual possess more than half of the direct psychological connections to an earlier individual that usually hold day to day in normal persons. This would avoid Locke’s worry: Cooper would not be morally responsible for an action of Audrey’s if he only q-remembered one action of hers because then Cooper would only be weakly psychologically connected to Audrey. On this view, moral responsibility requires that Cooper, to be morally responsible for any of Audrey’s actions, would have to be strongly psychologically connected to Audrey, and that requires more than one psychological connection.

Strong psychological connectedness seems like a promising contender for the DOC. Like mere psychological connectedness, it is non-transitive and it can hold one-many, so it avoids the counter-examples to non-branching psychological continuity (Bert’s Brain) and mere psychological continuity (Clive-1000). And it avoids the problem of loose psychological connections.

---

98 Indeed, it seems methodologically inconsistent of Locke to claim this. Recall that Locke argues that an individual would not be morally responsible for Socrates’ actions if she had Socrates’ soul to support his view that sameness of soul does not constitute personal identity. A defender of a soul view of personal identity could claim that God would stop such a thing happening in an attempt to undercut Locke’s point. The fact that Locke implicitly dismisses such a response undercuts his own use of that response to the problem of loose psychological connections.
The criterion of diachronic ownership according to the strong psychological connectedness view is the following:

An individual at t₂ diachronically owns (and hence is morally responsible for) a free action A performed at t₁ if and only if the individual at t₂ is strongly psychologically connected to the individual who A-ed at t₁.

While promising, this criterion is problematic because it implies an arbitrariness between being and not being morally responsible for some past action. Consider the following case, which is an adaption of Parfit’s (1984: 229-241) psychological spectrum case. John sits in a chair attached to a scientist’s ‘brain-machine’. If the scientist turns the dial a little bit, it will replace a small amount of John’s psychology with that of Mary’s. If the scientist turns the dial all the way, it will replace John’s psychology entirely with Mary’s. In the former case, it seems that John survives. In latter case, it seems that John does not survive – he has been replaced by Mary (or an individual who has exactly the same psychology as Mary). But what happens when the scientist turns the dial to around the half way point? Suppose the resulting individual has 49% of John’s psychology and 51% of Mary’s psychology. Parfit claims that there is no fact of the matter about whether John or Mary survives because identity is indeterminate. All we can say is that the resulting individual has 49% of John’s psychology and 51% of Mary’s psychology.

It would take me too far afield to evaluate Parfit’s claim that identity is indeterminate. I rather want to question whether the fused agent (call him or her ‘Fuse’) is morally responsible for the actions of Mary and John. If strong psychological connectedness is the DOC, Fuse is not

---

99 Strictly speaking, strong psychological connectedness isn’t half an individual’s psychology, but rather (as noted) half the number of connections that are normally preserved day to day. So, the resulting individual is 49% psychologically connected to John and 51% psychologically connected to Mary. It is easier to talk in terms of percentages of psychology rather than percentages of psychological connections – that is why I have chosen to talk that way. But bear in mind this only for ease of expression.

100 For an evaluation of the claim that identity is indeterminate see, for example, Noonan (2003).
morally responsible for John’s actions but he/she is morally responsible for Mary’s actions. Remember that strong psychological connectedness only holds when an individual at t₂ shares at least half (i.e. 50%) of the psychology of an individual at t₁. But it seems arbitrary that Fuse is not morally responsible for John’s actions but he is morally responsible for Mary’s. Notice that if the scientist had turned the dial a little more and Fuse had 50% of John’s and 50% of Mary’s respective psychologies, thenFuse would be morally responsible for both John’s and Mary’s actions. And if the scientist had turned the dial a little more and Fuse had 51% of John’s psychology but only 49% of Mary’s, then Fuse would be morally responsible for John’s actions but not Mary’s. It seems arbitrary for an individual to be or not to be morally responsible for past actions due to such a small difference in psychological states. It seems much more plausible that Fuse is morally responsible for at least some of both John’s and Mary’s actions in all the variations of this case. Therefore, strong psychological connectedness is not the DOC.

One way to deal with this is to posit some sort of appropriateness condition on diachronic ownership. It matters how an individual at t₂ is psychologically connected to an individual at t₁ – that is, it matters which psychological states an individual at t₂ shares (via a causal connection) with an individual at t₁. Fuse seems morally responsible for at least some of both John’s and Mary’s actions because he seems appropriately psychologically connected with respect to certain actions to both John and Mary. Which psychological states are relevant to diachronic ownership? Well, this depends on which psychological states we take to be relevant to synchronic ownership. The standard structuralist conditions offer three candidates: higher-order volitions, values, or mechanisms (though mechanisms aren’t strictly speaking psychological states – more on this momentarily). Let’s assume that values are the relevant psychological state. What

---

101 Another possibility is to think that diachronic ownership comes in degrees. I lack the space to evaluate this possibility adequately, though it seems to me that it will still require an appropriateness condition (or something further condition) because it would seem strange to say that Fuse is 49% responsible for all John’s actions and 51% responsible for all Mary’s actions when Fuse shares none of the states which lead to some of John’s and Mary’s respective actions, but shares all of the states which lead to some of John’s and Mary’s respective actions. It seems more plausible to say that Fuse is not morally responsible for actions he/she is not appropriately psychologically connected, whilst he/she is morally responsible for actions he/she is appropriately psychologically connected to.
I say about values should be directly applicable to higher-order volitions. However, my points will not apply directly to mechanisms. Fischer and Ravizza, unfortunately, never specify exactly what a mechanism amounts to; all they claim is that such mechanisms lead to action. Despite this, we might assume that mechanisms are underpinned by certain psychological states, since Fischer and Ravizza (1998) endorse the causal theory of action according to which actions are caused by psychological states. Thus if we knew which psychological states were important to mechanisms – that is, if we knew how to individuate mechanisms – then we could modify what I say about values so that it applies to the psychological states that underpin mechanisms.

Fuse, then, diachronically owns (and thus is morally responsible for) a free action \( A \) if and only if Fuse is appropriately value-connected to individual who performed \( A \), where ‘appropriate value’ means the values which were expressed by \( A \). Formally:

\[
\text{An individual at } t_2 \text{ diachronically owns (and thus is morally responsible for) a free action } A \text{ performed at } t_1 \text{ if and only if the individual at } t_2 \text{ shares value } V \text{ which was expressed by } A \text{ and is causally connected to the individual who } A\text{-ed at } t_1. 
\]

Call this criterion ‘appropriate value connectedness’. From this criterion it follows that Fuse is morally responsible for some of both John’s and Mary’s actions – namely, those actions which Fuse is appropriately value-connected to.

But this criterion does not avoid a version of the problem of loose psychological connections, which we might call the problem of loose value connections.\(^{102}\) (There are analogous problems for desires and mechanisms, so it is not a problem just for appropriate value connectedness, but rather for any DOC which focuses only on minimal elements of an individual’s psychology.) Suppose Audrey \( A\)-ed and in doing so expressed value \( V \).

---

\(^{102}\) Indeed, it has further problems. For example, an individual \( S \) could have acquired a value \( V \) expressed in action \( A \) via the individual who \( A\)-ed (\( R \)) because \( R \) persuaded \( S \) to endorse this value. I won’t respond to this problems, because I am ultimately not going to endorse this proposed DOC.
Neuroscientists then (using their advanced technology) isolate the neural substrate of that value and implant it into Cooper (whilst maintaining the causal connection between both individuals). Despite possessing $V$, Cooper does not seem morally responsible for $A$. But if appropriate value connectedness is the DOC, then he is morally responsible for $A$; hence, appropriate value connectedness is not the DOC.

Perhaps, then, what DOC requires is that an individual has the same (or mostly the same) set of values at later time. In other words, an individual at $t_2$ is morally responsible for a free action $A$ performed at $t_1$ if and only if the individual at $t_2$ shares the system of values and is causally connected to the individual who $A$-ed. The trouble with such a criterion, however, is specifying what a ‘system of values’ amounts to. Is it all of an individual’s values? It seems it can’t be because it seems implausible that individual might not be morally responsible for an action $A$ which expressed value $V_1$ just because they no longer possess value $V_2$. We therefore need some principled way to explain what counts as enough values (and perhaps other psychological states) for diachronic ownership. It seems hard to do this without appealing to some further consideration.

One possibility, perhaps, is to appeal to narrative views of persons. Such views include a narrative condition, which I appropriated for my synchronic ownership condition (see §3.2). Such a view might explain what counts as enough values for diachronic ownership in a non-arbitrary fashion. In the next section, I consider whether narrative views can supply the DOC, but conclude they cannot. I then develop a narrative-based DOC. As we’ll see, this DOC avoids explicitly appealing to specific psychological states, such as values.

4.7 Narrative connectedness

According to a narrative view of persons, an individual is a person if she conceives of herself as a protagonist in an on-going self-told story. I won’t assess whether such views are adequate accounts of personhood or personal identity. Narrative views do seem to supply a promising
DOC. Consider the Audrey/Cooper case. Intuitively, Cooper is not morally responsible for Audrey’s action. Narrative views accommodate this judgement: Cooper is not morally responsible for Audrey’s action because her action does not cohere with his narrative. In other words, actions belong to an individual if and only if they are part of her self-narrative – that is, her self-told story of her life. However, standard narrative views fall short because an individual’s self-narrative is a story of her entire life, and so seems to presuppose personal/numerical identity. This effectively rules out a life-narrative criterion as the DOC.

But, just because narrative views fall short, that doesn’t mean that some narrative condition might not do the trick. In §3.2, I appealed to narrative explanation as part of my authenticity condition on synchronic ownership. Narrative explanations, thus conceived, are not stories of an individual’s entire life; they are story-like explanations of an individual’s particular actions. Following Schroer and Schroer (forthcoming: 17-18), I contend that such stories do not extend to events that were experienced by time-slices of the individual to whom the individual is not now psychologically connected. Hence narrative explanations can play a role in the DOC.

In §3.2, I claimed that an individual synchronically owns an action $A$ if and only if $A$ is a free* action (that is, the individual satisfies one of the standard structuralist conditions on synchronic ownership when she $A$s) and the individual is able to provide a narrative explanation of $A$. I claimed that an individual is able to provide such an explanation – or, rather, forms the disposition to provide such an explanation – if she has a coherent or harmonious psychology (or psychological structure) when she $A$s.

Just as narrative explanations are partly constitutive of synchronic ownership, I claim that narrative explanations are partly constitutive of diachronic ownership. The main difference is that whereas a narrative explanation for synchronous ownership must make sense of $A$ with respect to the character, past actions, and/or planned future actions of an individual at the time of

---

103 This is implied by Schechtman (1996) and is made explicit by DeGrazia (2005).
action \( (t_1) \), a narrative explanation for diachronic ownership must make sense of \( A \) – that is, an action performed at \( t_1 \) – with respect to the character, past actions, and/or planned future actions of an individual at a later time \( (t_2) \). The crucial difference here is the individual at \( t_1 \)’s character: this might have changed between \( t_1 \) and \( t_2 \). For example, suppose an individual \( S \) \( A \)s at \( t_1 \). Suppose \( S \) synchronically owns \( A \) – that is, she satisfies the standard structuralist conditions when she \( A \)s and she can, if asked, narratively explain \( A \). Does \( S \) diachronically own \( A \) at \( t_2 \)?

Well, that depends on whether \( S \) can provide a narrative explanation of \( A \) at \( t_2 \), and that depends on whether \( S \) instantiates the relevant dispositional property at \( t_2 \). If \( S \) has changed such that she has a very different psychology (that is, beliefs, desires, values, etc.) than she once had (perhaps she underwent a radical religious conversion), then she will be unable to provide a narrative explanation for \( A \). That is, \( A \) no longer makes sense with who she now is; it does not contribute to her self-conception, and therefore she doesn’t instantiate the relevant dispositional property.

But if her psychology is more or less the same, then she will be able to provide a narrative explanation for \( A \), and therefore she instantiates the relevant dispositional property.

While being able to provide a narrative explanation of some action \( A \) is necessary for diachronic ownership, it is not sufficient. It is possible that \( S \) could provide a narrative explanation for some past action \( A \) performed by \( R \) to whom \( S \) is not psychologically connected. For example, suppose Larry is an instant agent who was created with a fully developed psychology such that he believes he is a famous military commander (call him Napoleon) who commanded his army to victory many times. Let’s suppose that Larry has a qualitatively identical psychology to the real Napoleon, but that Larry is not psychologically connected to him. So Larry ‘remembers’ experiences qualitatively identical to actual experiences of Napoleon’s, but those ‘memories’ are not q-memories (which require a psychological connection). Because Larry has, by hypothesis, a fully developed psychology, he can narratively explain his post-creation actions and therefore may be morally responsible for those actions.
Larry will also be able to provide narrative explanations for those of Napoleon’s actions of which he has fake memories – that is, actions performed prior to Larry’s creation, which Larry seems to remember but to which he in fact bears no psychological connection at all. But it would be absurd to suppose that Larry is morally responsible for those actions. It also matters that the latter individual is properly connected to the earlier individual who acted freely. Because Larry is not properly connected to Napoleon, he is not morally responsible for Napoleon’s actions. Thus, diachronic ownership requires that an individual must be able to provide a genuine narrative explanation; this requires that the psychological features underpinning the disposition to provide the narrative explanation are q-states that were actually involved in the production of $A$ and are those q-states which actually made sense of $A$ (that is, those states which underpin the disposition to provide a narrative explanation of $A$ when it was performed). While Larry may be morally responsible for his post-creation actions, he is not morally responsible for any of Napoleon’s actions because he (Larry) is not by hypothesis q-related to Napoleon. If Larry possessed q-states that actually lead to Napoleon’s actions – perhaps Larry is a post-fission duplicate of Napoleon, or some such – and Larry is able to narratively explain Napoleon’s actions, then Larry may be morally responsible for those actions. The DOC of the structural-narrative view is therefore the following:

An individual at $t_2$ diachronically owns (and thus is morally responsible for) a free action $A$ performed at $t_1$ if and only if the individual at $t_2$ can provide a genuine narrative explanation of $A$.

Call this DOC ‘narrative connectedness’.104 I should note that I don’t intend this condition to be anything more than a DOC. It therefore shouldn’t be expected to fulfil all that might be required

---

104 Cf. Schroer and Schroer (forthcoming: 15).
of an ‘ownership’ condition of a narrative view of persons, such as underpinning other practical concerns. Standard narrative views are concerned with an individual’s life story, and so might include actions and events for which an individual was never and/or is no longer morally responsible. Thus, as noted above, such narrative accounts are not plausible candidates for the DOC. Narrative connectedness, on the other hand, is only concerned with moral ownership – that is, the actions for which we are morally responsible.

Narrative connectedness avoids the counter-examples which undermine all the other candidates for the DOC. Fuse – the post-fusion individual – is morally responsible for John’s or Mary’s actions in so far as he can provide genuine narrative explanations for those past free actions, and it seems plausible that there are circumstances in which Fuse is going to be able to provide genuine narrative explanations for some of those actions. For example, suppose John swore at an old man a few days before he was fused with Mary, and John synchronically owned that action. It seems possible that Fuse has enough of John’s beliefs, desires, values and memories to be able to provide a narrative explanation of John’s action. Fuse might say that John swore at the old man because he doesn’t like old men as they remind him of his abusive father, and this old man (unknowingly) took a stance that his father used to take just before he would beat him. All the beliefs, desires, values, and memories at play here would, of course, only be q-states because ‘normal’ states presuppose personal identity and we’ve assumed that Fuse is not identical to John. It seems that Fuse will also be able to provide genuine narrative explanations for some of Mary’s actions. There will, of course, be actions of Mary’s and John’s that Fuse will be unable to narratively explain. Suppose that Mary gave a large sum of money to charity a few years before. If Fuse lacks the relevant (q-)beliefs, desires, values, and memories related to this action, then Fuse is not going to be able to provide a narrative explanation of Mary’s action, and thus Fuse will not diachronically own Mary’s action. Hence, narrative connectedness does not imply an arbitrary difference between being and not being morally responsible for past free actions.
Narrative connectedness also avoids the problem of loose value/psychological connections. I'll only consider the problem of loose value connections, though my response should generalise to other psychological connections. Suppose Cooper is merely implanted with a value $V$ of Audrey’s that she expressed in some past free action $A$. Cooper seems not to be morally responsible for $A$. While Cooper is by hypothesis psychologically (i.e. value) connected to Audrey, Cooper will not be able to provide a narrative explanation for $A$. Why? Well, he lacks Audrey’s other beliefs, desires, values and memories, so he will have no inkling about how to even start providing a narrative explanation of $A$. In other words, because he lacks the underlying psychological elements, Cooper lacks the disposition to provide a narrative explanation of $A$. But what if Cooper is somehow able to confabulate a story about why ‘he’ $A$-ed? (This might involve stipulating that Cooper is the sort of person who $A$’s, but didn’t $A$ when Audrey $A$-ed, and that $A$ manages to make perfect sense with respect to Cooper’s other actions.) While Cooper might be able to make up some story, it won’t be a genuine narrative explanation because Cooper lacks (a) the q-states involved in the production of $A$ and (b) the q-states that made sense of $A$ (that is, the states that would underpin the relevant dispositional property) when it was performed.

Of course, we might be able to imagine (a) being satisfied (assuming that the quantity of q-states involved in the production of an action is quite low). But can we imagine (b) being satisfied? Yes: if Cooper is implanted with enough of Audrey’s other psychological states, there must be a point at which Cooper can provide a genuine narrative explanation. But I contend that for Cooper to articulate a sufficiently intelligible narrative explanation of $A$, and thus to instantiate the relevant dispositional property, he will have to be implanted with so many of Audrey’s psychological states that this case turns into a fusion case. That is, the resulting individual will possess large swathes of Audrey’s and Cooper’s psychological states. The resulting individual might be able to provide a genuine narrative explanation of $A$ because he or she will be a fusion of Cooper and Audrey. But that is an acceptable result for the same reason that it’s acceptable
that Fuse may be morally responsible for some of both John’s and Mary’s action. So, my structural-narrative view avoids the problem of loose value/psychological connections.

Narrative connectedness is also a non-transitive relation and it can also hold one-many, so it avoids the fission counter-examples to non-branching psychological continuity and the Clive-1000 counter-example to mere psychological continuity.

The DOC of my structural-narrative view, then, is narrative connectedness. As I argued in §4.3, the DOC determines the locus of moral responsibility. We are inclined to hold that the ‘whole’ temporally extended individual is the locus of moral responsibility because we assume numerical identity is the DOC. However, because the DOC is narrative connectedness, it follows that the locus of moral responsibility, on the structural-narrative view, is something which is not as temporally extended as the ‘whole’ individual; I propose that it is the narrative self.

One way to think of narrative selves is as a collection of time-slices of individuals that are united by narrative connectedness. So a narrative self is a collection of time-slices (or person-stages) of an individual for which: the individual (at a particular time) can provide genuine narrative explanations for the actions of those time-slices. On this view, the time-slice who A-ed is initially morally responsible for A-ing, and then moral responsibility for A-ing is transmitted to other time-slices who are narratively connected to the time-slice who A-ed. Being a particular narrative self is thus a convenient way of talking about the transmission of moral responsibility from one time-slice to another.\(^{105}\)

While there is certainly more work to be done to develop the structural-narrative view, I have provided enough so that I can now make clear that Brainwashed Beth is not a counter-example to the structural-narrative view.

\(^{105}\) Another proposal comes from Shoemaker (1999) and Strawson (2011). According to them, the ‘self’ (which is similar to the narrative self), is perspectival. This means that the boundaries of the self are not as clear cut as I have been suggesting, and rather depend on the perspective of the individual at a particular time. This is a somewhat complicated proposal, and for the sake of simplicity I have not discussed it here.
4.8 Back to Brainwashed Beth

It’s now time to return to Brainwashed Beth – the control case that purports to be a counter-example to the structuralist conditions on moral responsibility, including my structural-narrative view. In this case Beth, an exceptionally sweet woman, is brainwashed to have the system of values of a notorious serial killer, Chuck. As a direct consequence of being given such a system of values, Beth kills her neighbour, George. It seems that Beth is not morally responsible for killing George, yet she seemingly satisfies the structuralist conditions on moral responsibility. Structuralism therefore seems to be false.

Historicists claim that cases like Brainwashed Beth show that there are historical conditions on synchronic ownership; thus historicists claim that we judge that Beth is not morally responsible because she does not satisfy some historical condition on synchronic ownership. But compatibilists cannot be historicists. As I argued in §2.7, historicism collapses into source incompatibilism, so the historicist’s diagnosis of Beth’s non-responsibility actually speaks in favour of source incompatibilism. Compatibilists therefore have to establish an alternative diagnosis to the historicist’s.

In Chapter 2 I argued that there are two reasons that we might judge that Beth is not morally responsible. The first is that it seems unfair that the exceptionally sweet Beth will be morally responsible for the nasty actions of her brainwashed self. The second is that Beth is covertly controlled by the neuroscientists. Both of these reasons can be used to motivate historicism, so both reasons really just motivate source incompatibilism. In this section, I will undercut both of these reasons, thereby showing that we have no reason to think that Beth is not morally responsible for killing George. I’ll start with the first reason before moving onto the second.

---

106 In §3.4, I proposed a third reason: we presume that Beth, because she has been brainwashed to have the values of a serial while she has the beliefs and memories of a sweet and kind person, is not a coherent agent when she kills George. I’ve set aside this third reason because I’ve assumed, as noted in §3.4, that Beth is a coherent agent when she acts.
In this chapter I have argued that the locus of moral responsibility is not the whole temporally extended individual but rather the narrative self, since the DOC is narrative connectedness. I will now argue that the nasty Beth is a different narrative self to the nice Beth, so there is no transfer of moral responsibility between these two distinct narrative selves; hence it is fine to accept that the nasty Beth is morally responsible.

Consider first another feature of the Brainwashed Beth case that I have ignored so far. In order to support his claim that Beth is not morally responsible, Mele (2009b: 471) asks us to suppose that a day after George is killed, Beth regains pre-brainwashing Beth’s (BethV₁’s) system of values – that is, the values of an exceptionally sweet person. So for most of her life Beth – that is, the whole, temporally extended individual – has been sweet and innocent, apart from that day when she decided, in line with her values of the day, that she should kill George. Given this, I contend that BethV₁ and post-brainwashing Beth (BethV₂) are different narrative selves – that is to say, BethV₁ (when she returns) is not narratively connected to BethV₂. To establish whether BethV₂ is a different narrative self to BethV₁, we must ascertain whether BethV₁ (when she returns) is able to provide a narrative explanation of BethV₂’s action of killing George.

While it seems that BethV₂ is able to provide a narrative explanation for her action of killing George, it seems that BethV₁ (when she returns) will be unable to do so. BethV₁ has none of the values, beliefs, or desires of serial killer; at most, she only remembers BethV₂’s heinous actions. Recall, as I argued in §3.2, that if an individual has a coherent or harmonious psychological structure at the time of action, then she forms the disposition to provide a narrative explanation of her action when asked to. So, while BethV₂ has such a disposition, it is underpinned by her psychology at the time of action. Given that BethV₁ has a different psychology to BethV₂ – in particular, BethV₁ has the values of an exceptionally sweet person, and BethV₂ has the values of a serial killer – there is nothing about BethV₁’s psychology that will

107 Mele is actually discussing a slightly difference case, but he says that this detail can be applied to Brainwashed Beth.
underpin a disposition to provide a narrative explanation of Beth\textsubscript{V2}’s action; hence Beth\textsubscript{V1} will be unable to provide a narrative explanation of Beth\textsubscript{V2}’s action of killing George. It seems clear, then, that Beth\textsubscript{V1} and Beth\textsubscript{V2} are not the same narrative self: Beth\textsubscript{V1} has the system of values of a sweet and innocent person and Beth\textsubscript{V2} has the system of values of a serial killer; Beth\textsubscript{V1} is simply unable to provide narrative explanations of Beth\textsubscript{V2}’s actions.

It’s not surprising that many judge that Beth is not morally responsible, because Beth’s dominant narrative self (that is, Beth\textsubscript{V1}) – that exists for almost all of Beth’s existence, aside from the 24-hour blip when Beth (qua individual) murders George – is clearly not morally responsible for killing him. However, the important question is whether Beth\textsubscript{V2} – that is, the time-slice of Beth who exists for that brief 24-hour period – is morally responsible. Because we can distinguish between Beth\textsubscript{V1} and Beth\textsubscript{V2} (qua narrative selves), then the question ‘is Beth morally responsible?’ is, as it stands, not well formed because it assumes that there is one locus of moral responsibility throughout the pre- and post-brainwashing periods, namely Beth (the individual). That is to say, it assumes that the orthodox view, that personal identity is the DOC and thus the whole temporally extended individual is the locus of moral responsibility, is true. But the orthodox view is false and the whole individual is not the locus of moral responsibility, so this question is badly formed; the locus of moral responsibility is in fact the narrative self, which is a collection of narratively connected time-slices. Hence, the question we should ask – the question that is well formed – is a question about the moral responsibility of a narrative self. And of course there are two different questions we might ask here: one about Beth\textsubscript{V1}’s, and one about Beth\textsubscript{V2}’s, moral responsibility.

If we ask the question about Beth\textsubscript{V1}, then we can only seriously be asking about diachronic ownership, since Beth\textsubscript{V1} wasn’t even in existence at the time of the murder and hence cannot possibly synchronically own it. And Beth\textsubscript{V1} does not diachronically own the murder: since Beth\textsubscript{V1} and Beth\textsubscript{V2} are different narrative selves, Beth\textsubscript{V2}’s moral responsibility (if any) for killing George does not transfer to Beth\textsubscript{V1} when she is restored. And this result does nothing to
motivate historicism: the historicist’s claim that there must be historical conditions on moral responsibility is a claim about synchronic ownership: it is a claim about non-responsibility at the time of action. And, as we have just seen, we don’t need historicism in order to get the (right) result that BethV1 does not synchronically own, and hence is not morally responsible for, the murder.

If, on the other hand, we ask the question about BethV2, it can only sensibly be a question about synchronic ownership (since by stipulation BethV2 is a single narrative self, and so she, post-murder, will be morally responsible for the murder if and only if she was morally responsible for it at the time of the murder). So the question we need to answer is: does BethV2 synchronically own, and hence is she morally responsible for, the murder?

While the relevant question is about BethV2, one reason, as I’ve noted, for judging that BethV2 is not morally responsible is that it is assumed that BethV2’s responsibility for killing George will somehow be transferred to BethV1 when she returns. It is crucial, then, that BethV1 and BethV2 are different narrative selves, and it seems clear, as I’ve argued, that they are. Given that BethV1 and BethV2 are different narrative selves, there is no transmission of moral responsibility between them. So if BethV2 is morally responsible for something, that doesn’t mean that BethV1 also must be. Hence, if BethV2 is morally responsible for killing George, her moral responsibility for that action will not transmit to BethV1 when she returns the day after George is murdered. Even if BethV1 never returned, it is not the case that responsibility for killing George would be retrospectively transferred onto BethV1, who existed prior to George being killed. BethV1 goes out of existence as soon as Beth is brainwashed to have the values of a serial killer, and a new narrative self – BethV2 – comes into existence. So, again, if BethV2 is morally responsible for killing George, BethV1 will not get the blame for killing George. BethV1’s slate remains untainted by BethV2’s action no matter what happens.

We are now in a position to ask the correct question: is BethV2 morally responsible for killing George? There no longer seems to be much intuitive pull towards thinking that she is not.
The worry with holding that Beth\textsubscript{2} is morally responsible is that Beth\textsubscript{1} will be blameworthy for Beth\textsubscript{2}’s actions. At root, this is a worry about the transmission of moral responsibility from one time-slice to another. However, my structural-narrative view does not have this consequence. The fact that Beth\textsubscript{2} is morally responsible for killing George does not entail that the sweet and innocent Beth\textsubscript{1} will get the blame. So, one reason for judging that Beth is not morally responsible has been undermined.

Critics of structuralism might now claim that Beth\textsubscript{2} is not morally responsible because she has been \textit{covertly controlled} by her manipulators. They might even insist that this – and not the illicit assumption that the whole extended person is the locus of moral responsibility – is the main reason that we judge that Beth\textsubscript{2} is not morally responsible. Structuralists, however, can also undercut this reason; they can do so by employing a modified version of Fischer’s response to Designed Ernie/Designed Ernie*.

Fischer (2011), as I noted in §2.7, argues that the judgement that Ernie* (or Ernie) is not morally responsible is unreliable because the opposing judgement can be straightforwardly generated. Here is a version of Fischer’s argument. If Designed Ernie* is a counter-example to all plausible current compatibilist conditions on moral responsibility, then there must be no differences relevant to moral responsibility between Ernie* and a merely causally determined counterpart of Ernie* (call him Bernie). Now if Bernie has not been covertly controlled, then we have no reason to think he is not morally responsible; so we should assume he is morally responsible. But if there are no relevant differences between Bernie and Ernie*, it follows that Ernie* is morally responsible. Of course, we could run the argument the other way and conclude that Bernie is not morally responsible. But the fact that the opposing judgement to one required to get this argument off the ground can so easily be generated by changing the order of the cases shows that the non-responsibility judgement about Ernie* is not reliable. So Designed Ernie* is not a counter-example to all plausible current compatibilist conditions on moral responsibility;
hence a manipulation argument motivated by a design case fails. Call Fischer's move the ‘reverse generalisation strategy’.108

Structuralists can use the reverse generalisation strategy in response to brainwashing cases to undercut the reliability of the non-responsibility judgement about Beth\textsubscript{V2}. According to my structural-narrative view, the locus of moral responsibility is the narrative self – that is, a time-slice (or a collection of time-slices) of the individual. Now consider a merely causally determined counterpart of Beth\textsubscript{V2} called Ruth. She has lived out her entire life without any sort of untoward manipulation or covert control by other agents but is otherwise psychologically exactly similar to Beth\textsubscript{V2}. (Of course, Ruth’s psychology is a bit weird: while she has the values of a serial killer, she, like Beth, remembers being a nice person too. Again, we’re assuming that Ruth, like Beth, is a coherent agent when she acts so that she, like Beth, satisfies my proposed synchronic ownership conditions.) Ruth at \( t_1 \) shares Beth\textsubscript{V2}’s precise psychological profile at the time at which she murders George – that is, in time-slice terms, Beth\textsubscript{V2} and Ruth are exactly similar. And like Beth, Ruth kills her neighbour. As Ruth has not been covertly controlled, we have no reason to think that she is not morally responsible; so we should assume she is morally responsible for murdering her neighbour. If Brainwashed Beth is a counter-example to all structuralist conditions on moral responsibility, there must be no differences relevant to moral responsibility between Beth\textsubscript{V2} and Ruth at \( t_1 \). But if Ruth is morally responsible, then it follows that Beth\textsubscript{V2} is morally responsible too. Again, we can easily elicit the opposing judgement to one required to motivate an argument against structuralism. Therefore, the non-responsibility judgement about Beth\textsubscript{V2} is unreliable. Hence, Brainwashed Beth is not a counter-example to the structural-narrative view. This result can be generalised to all brainwashing cases as they all have a similar structure to Brainwashed Beth.

108 In Chapters 6 and 7, I call the sort of reply that uses the reverse generalisation strategy a ‘medium’ compatibilist reply to distinguish it from soft and hard compatibilist replies, respectively.
4.9 Conclusion

In Part One (Chapters 2-4), I have sought to defend structuralism – the view that synchronic ownership depends only on facts about an individual’s psychology at the time of action – from hypnosis and brainwashing control cases. These control cases are used by historicists to motivate their view. In Chapter 2, I first argued that, due to thought experiments inspired by those in the personal identity literature, control cases do not provide decisive motivation for historicism. I then argued that historicism is methodologically inconsistent, and that removing this inconsistency leads to historicism collapsing into source incompatibilism. Hence historicism is untenable.

Despite historicism being untenable, I admitted that there are some intuitively plausible features of historicism. In Chapter 3, I aimed to capture one of these features. I conceded – due to hypnosis cases – that the standard structural conditions (e.g. Frankfurt’s hierarchical conditions) were insufficient for synchronic ownership, and therefore needed to be supplemented with an authenticity condition. This is what historicists hold too, but I claimed that the historicist’s error is the presumption that authenticity is a historical matter. I instead argued that there is a plausible structural authenticity condition and I proposed that the standard structuralist conditions should be supplemented with such a condition – namely the requirement that synchronic ownership requires that an individual be able to provide a narrative explanation for her action. I argued that hypnotised individuals are unable to satisfy such a condition, and thus hypnosis cases are not counter-examples to my structural-narrative view. I then argued that brainwashed individuals probably don’t satisfy my authenticity condition either, but I conceded that it might be conceptually possible for a brainwashed individual to satisfy it.

In this chapter, I have shown that brainwashing cases are not counter-examples to my structural-narrative view. In doing so, I also aimed to capture another intuitively plausible feature of historicism – namely, that it explains why it might seem that Chuck is more responsible than Beth$_{v2}$. This is because Chuck, unlike Beth$_{v2}$, is morally responsible for becoming the horrible
person he now is. Historicists, I claimed, mistakenly conclude that this means that Beth\textsubscript{V2} is \textit{not} morally responsible. I then set out to argue that we can plausibly hold that Beth\textsubscript{V2} \textit{is} morally responsible. I first argued that the locus of moral responsibility is not the whole temporally extended individual, but rather the narrative self. This allowed me to show that one reason for judging that Beth\textsubscript{V2} is not morally responsible – viz. that Beth\textsubscript{V1} will get the blame for Beth\textsubscript{V2}'s actions – does not hold, because, according to my structural-narrative view, responsibility for killing George (Beth’s neighbour) does not transfer from Beth\textsubscript{V2} to Beth\textsubscript{V1} (when she returns a day later, or retrospectively). Finally, I argued that another reason to judge that Beth\textsubscript{V2} is not morally responsible – viz. that she has been covertly controlled – does not hold because this judgement can be undercut my emphasising that there are no relevant differences between Beth\textsubscript{V2} and a merely causally determined counterpart time-slice of hers.

This ends Part One, and my defence of structuralism in the debate between compatibilists. In Part Two, I turn my attention towards the debate between compatibilists and incompatibilists. In Chapter 5, I focus on structural issues about manipulation arguments, and I argue that compatibilists’ efforts should be focused on showing why control cases are not in fact counter-examples to all plausible current compatibilist conditions on moral responsibility. In Chapters 6 and 7, I argue that there are no such counter-examples.
PART TWO

COMPATIBILISM AND MANIPULATION ARGUMENTS
Chapter 5: Manipulation Arguments against Compatibilism

5.1 Introduction

Manipulation arguments claim to show that compatibilism is false. What I shall call ‘the manipulation argument-scheme’ (or MAS) runs as follows:

M1. (Judgement) An individual covertly controlled in manner \( X \) to perform action \( A \) is not morally responsible for \( A-ing \).

M2. (No difference) There are no differences relevant to moral responsibility between an individual who is covertly controlled in manner \( X \) to \( A \) and an individual who is causally determined to \( A \).

Therefore,

M3. If causal determinism is true, then no one is morally responsible for anything they do. In other words, compatibilism is false.

Different instances of MAS are supported by different control cases; thus M1 of a particular instance of MAS is supported by a judgement about a particular control case, such as Brainwashed Beth. When faced with an instance of MAS, compatibilists therefore must either reject M1 or M2.\(^{109}\) In order to reject M1, a compatibilist must argue that a particular covertly controlled individual \emph{is} morally responsible. In order to reject M2, a compatibilist must argue that there is some relevant difference between a covertly controlled individual of the relevant kind and a merely determined one, and this usually (though not exclusively, as we'll see in §5.5) involves positing a further condition on moral responsibility.

Although my principal aim in Part One was to settle an in-house dispute between compatibilists, I also implicitly showed that three manipulation arguments – that is, instances of

\(^{109}\) Cf. McKenna (2008a: 143)
MAS – are unsuccessful. I showed that a manipulation argument supported by a hypnosis case is unsound because there is a relevant difference between a hypnotised individual and a merely determined individual – namely, a merely determined individual can provide a narrative explanation of her actions, while a hypnotised individual cannot. In other words, I showed that M2 of a manipulation argument supported by a hypnosis case is false. In response to manipulation arguments supported by brainwashing cases, such as Brainwashed Beth, I showed that manipulation arguments supported by such cases are unsound by arguing that the judgement that Beth is not morally responsible is without support if it is based on the assumption that the ‘innocent’ time-slices of Beth will get the blame for brainwashed time-slices of Beth’s actions. And I argued the non-responsibility judgement about Beth is unreliable if it is based on the fact brainwashed Beth is covertly controlled. In other words, I argued that M1 of this manipulation argument is false, and so manipulation arguments supported by such cases do not get off the ground. I also implicitly argued that manipulation arguments supported by design cases are unsound by arguing that our judgement about designed individuals is unreliable; hence M1 of such arguments is false.

In Part Two, I turn my attention to the remaining manipulation arguments and the control cases that support them. In this chapter I focus on clarifying the mechanics of manipulation arguments. Among other things, I argue that manipulation arguments to the best explanation – that is, those that include a ‘best explanation’ of our judgements about covertly controlled individuals – collapse into ‘straight’ manipulation arguments (that is, instances of MAS). I also argue that both sorts of argument succumb to the ‘no generalisation’ objection – which claims we cannot generalise our judgements about covertly controlled individuals to merely determined individuals; hence M2 is false. I then sketch two ways for incompatibilists to avoid this objection, which sets the scene for my arguments in subsequent chapters.

In Chapters 6 and 7, I turn my attention towards defending compatibilism. I argue that none of the remaining control cases (the cases I shall consider feature sporadic and non-sporadic
continuous manipulation, ego button manipulation, programming, and conditioning) support an argument against compatibilism. In Chapter 6, I argue that there is an ambiguity over who the locus of moral responsibility is in sporadic and non-sporadic continuous manipulation cases, and I argue that once we clarify who the locus of moral responsibility is in these cases it becomes clear that these cases are not counter-examples. In Chapter 7, in response to the other cases (the ego button, programming, and conditioning), I present three related arguments. The resulting position is (a form of) hard compatibilism – the view that certain covertly controlled and causally determined individuals are, in fact, morally responsible.

The structure of this chapter is as follows. In §5.2, I outline Pereboom’s (2001) Four-Case Argument, which features three control cases (Cases 1-3) and a mere determination case (Case 4). In §5.3 I show that Pereboom includes a ‘best-explanation’ premise – which purports to identify the source of our judgements about covertly controlled individuals – within his Four-Case Argument, and thus considers his argument to be a best-explanation manipulation argument, rather than a straight manipulation argument. In §5.4 I argue, contra Pereboom and (to an extent) following Mele, that covert control is in fact the best explanation for the judgement that Plums 1-3 are not morally responsible. In §5.5 I argue that both straight and best-explanation manipulation arguments succumb to the ‘no generalisation’ objection unless the conclusion of those arguments is modified. In §5.6 I argue that the inference to the best explanation is not required to secure the conclusion that compatibilism is false. I show that support for this premise is the truth of the ‘no difference’ premise, but that the truth of the ‘no difference’ premise (in conjunction with the judgement premise) is sufficient to show that compatibilism is false; so there’s no need to include a ‘best-explanation’ premise in a manipulation argument. In §5.7 I provide incompatibilists with another way to avoid the ‘no generalisation’ objection: forgo the ‘no difference’ premise, and argue that compatibilism is false purely on the basis of there being a control counter-example to all plausible current compatibilist
conditions on moral responsibility. To save compatibilism, then, we must show that no control case is in fact a counter-example to all these conditions. That is the task of Chapters 6 and 7.

5.2 The Four-Case Argument

In order to understand the intricacies of manipulation arguments, I will focus on Pereboom’s (2001) Four-Case Argument, as it is the most prominent and most widely discussed of all the manipulation arguments. I will present it in more or the less the fashion that Pereboom originally presented it.

The Four-Case Argument aims to show that compatibilism is false. To reach this conclusion, it features four cases: three control cases and one case of mere causal determination. The thought is, I take it, that we should assume the target of the argument initially takes herself to have no reason to believe that causal determinism precludes moral responsibility, and perhaps even believes that moral responsibility is compatible with causal determinism. The first of the cases (Case 1) is therefore (seemingly) far removed from a case of mere causal determination so that the target of the argument is not in a position to simply assume that the covertly controlled individual is morally responsible. And, Pereboom thinks, the target of the argument should judge that the covertly controlled individual is not morally responsible. The catch is that it seems that the covertly controlled individual also satisfies four of the leading sets of compatibilist conditions on moral responsibility: Frankfurt’s (1971) hierarchical conditions, Fischer and Ravizza’s (1998) moderate reasons-responsiveness condition, Hume’s (1739/1979: 399-411) character condition, and Wallace’s (1994) moral sensitivity condition. Thus Case 1 seems to be a counter-example

---

110 On Frankfurt’s view, an individual is morally responsible only if the structure of her will is in order (i.e. her effective first-order desires conform to her higher-order volitions). On Fischer and Ravizza’s view, an individual is morally responsible only if her action-producing mechanisms are moderately reasons-responsive (i.e. those mechanisms are regularly receptive to reasons and weakly reactive to them). On Hume’s view, an individual is morally responsible only if her actions stem from her own character. On Wallace’s view, an individual is morally responsible only if she is sensitive to moral reasons. These, of course, are only necessary conditions on moral responsibility: But Pereboom (2001: 111) claims that we can assume that the Plums satisfy all the other non-controversial conditions (at least in this context) on moral responsibility (i.e. synchronic ownership), such as epistemic conditions.
to all these conditions. Pereboom claims that Case 1 might be enough to ‘convince some compatibilists to abandon their position’ (2001: 112). But he notes it might not convince them all. The point of the remaining cases, according to Pereboom, is to show that ‘an agent’s non-responsibility under covert manipulation generalizes to the ordinary situation’ (2001: 112), and if non-responsibility generalises from a covertly manipulated/controlled individual to a merely determined individual, then this would force an incompatibilist conclusion.

Pereboom’s next cases (Cases 2 and 3) get progressively more like a case of mere causal determination (Case 4). The important point is that there are apparently no relevant differences between any of the three control cases and the case of mere causal determination. So the idea is that if compatibilists (or neutral inquirers112) accept that the individual in the first case is not morally responsible, they will be forced to admit that the individuals in other cases are also not morally responsible. And, of course, this means that they will have conceded that a causally determined individual is not morally responsible no matter what conditions she satisfies, and hence that incompatibilism is true.

Let’s start with Pereboom’s first case:

Case 1: Professor Plum [referred to hereafter as ‘Plum1’] was created by neuroscientists, who can manipulate him directly through the use of radio-like technology, but he is as much like an ordinary human being as possible, given his history. Suppose these neuroscientists “locally” manipulate him to undertake the process of reasoning by which his desires are brought about and modified – directly producing his every state from moment to moment. The neuroscientists manipulate him by, among other things, pushing a series of buttons just before he begins to

111 Notably, though, Pereboom doesn’t include historicist conditions among these compatibilist conditions. He does, however, argue that Plum1 satisfies these conditions after he sets out his four cases – see Pereboom (2001: 120-122).
112 These are participants in the debate who are committed to neither compatibilism nor incompatibilism, but who are open to each position (Pereboom 2008a: 162).
reason about his situation, thereby causing his reasoning process to be rationally egoistic. Plum is not constrained to act in the sense that he does not act because of an irresistible desire – the neuroscientists do not provide him with an irresistible desire – and he does not think and act contrary to character since he is often manipulated to be rationally egoistic. His effective first-order desire to kill Ms. White conforms to his second-order desires. Plum’s reasoning processes exemplifies the various components of moderate reasons-responsiveness. He is receptive to the relevant pattern of reasons, and his reasoning processes would have resulted in different choices in some situations in which the egoistic reasons were otherwise. At the same time, he is not exclusively rationally egoistic since he will typically regulate his behaviour by moral reasons when the egoistic reasons are relatively weak – weaker than they are in the current situation. (Pereboom 2001: 112-113)

Is Plum1 morally responsible for killing White? Plum1 has been covertly controlled via what I shall call ‘sporadic continuous manipulation’. It is sporadic because he is not always covertly controlled (‘he is often [that is, not always] manipulated to be rationally egoistic’), and it is continuous because when is manipulated they ‘directly produce his every state from moment to moment’. As a result of initiating and regulating his reasoning processes, the neuroscientists directly control Plum1’s actions whenever they press their buttons. Given that he kills White because he is under the covert control of the neuroscientists, it seems that Plum1 is not morally responsible for killing White. Indeed, it seems that Plum1 satisfies the leading compatibilist conditions, which I discussed above.

With this (alleged) counter-example in hand, Pereboom presents a second case, which he claims is not relevantly different from his first case:
Case 2: Plum [hereafter Plum2] 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 behavior 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. (Pereboom 2001: 113-114)

Is Plum2 morally responsible for killing White? Plum2 has been *programmed* to be just as rationally egoistic as Plum1. Given that Plum1 is not morally responsible and that there are supposed to be no relevant differences between Case 1 and Case 2 – that is, Plum2 seems to be under the neuroscientists’ control just as much as Plum1 is – it seems that Plum2 is not morally responsible. Again, it seems that Plum2 satisfies the leading compatibilist conditions on moral responsibility. Hence, it seems that Case 2 is also a counter-example to those conditions.113

A critic might be tempted to dispute Pereboom’s claim that there are no relevant differences between Case 1 and Case 2. The critic might concede that it is clear that Plum1 is not morally responsible because he has been manipulated from moment to moment and so is clearly

113 Todd (2011) argues that the incompatibilists can weaken their dialectical load (so to speak) by simply claiming that manipulation (or, more precisely, covert control) *mitigates* rather than undermines an individual’s moral responsibility, and he uses Pereboom’s Case 2 as his example. Todd argues that compatibilists cannot adequately explain this mitigation, and so compatibilism is false. Since I will argue in Chapters 6 and 7 that control cases are not counter-examples to all plausible current compatibilist conditions on moral responsibility, my argument will *a fortiori* undercut Todd’s argument. While I lack the space to properly assess Todd’s argument, I should note that my main point of contention with Todd is that it is not clear that compatibilists or even neutral inquirers should accept his claim that Plum2’s programming mitigates his moral responsibility for killing White. The only reason we might think that is if we think that programming is in some way different from being causally determined. Given that Todd claims there are no relevant differences between them, compatibilists and neutral inquirers should hold that Plum2 is morally responsible. I develop my version of this reply in Chapter 6, though I do not extend it explicitly to Todd’s argument. For direct replies to Todd see Tierney (2013), Capes (2013), and Khoury (2014).
controlled by the neuroscientists, but they might contend that it is far from clear that Plum2 is controlled by the neuroscientists, and so it’s not clear that we should judge that Plum2 is not morally responsible. The critic could even point out that Pereboom admits the neuroscientists do not directly control Plum2 like they do Plum1.

However, as Pereboom (2001: 114) says, ‘it would seem unprincipled to claim that here, by contrast with Case 1, Plum[2] is morally responsible because the length of time between the programming and the action is great enough.’ Indeed, sporadic continuous manipulation – as we have in Case 1 but not in Case 2 – is not the only way for a manipulator to control another agent; for example, as I argued in §2.7, a manipulator can still control an individual by designing her zygote in a particular way in conjunction with sufficient knowledge about the future (or local control of her environment). Likewise, the neuroscientists could control Plum2 by programming his neural states to be a particular way in conjunction with sufficient knowledge about the future (or by locally controlling his environment). Now, at least in this version of Case 2, Pereboom is not explicit that the neuroscientists have knowledge of Plum2’s future environment, though this detail can be added to the case without affecting its conceptual coherence.114 For Pereboom’s argument to go through, it must be irrelevant whether the manipulators’ control is direct or indirect; let’s grant that it is for the sake of argument. Given that there seem to be no relevant differences between Cases 1 and 2, then, it seems that the neuroscientists’ indirect control of Plum2 is just as responsibility-undermining as the neuroscientists’ direct control of Plum1.115

Pereboom’s next case moves much closer to a case of mere causal determination:

---

114 Pereboom (2005: 238) makes this move in response to a worry of McKenna’s. Pereboom also suggests that manipulators could have local control over Plum2’s environment. This is the sort of control that the production team have over Truman in The Truman Show. Carl Ginet sketched a similar sort of case as an objection to Fischer’s historicist view; see Fischer (2000: 414-415).

115 There are similarities between design cases and programming cases. Although I argued in Chapter 4 that design cases do not show that compatibilism is false, it seems that if Pereboom is correct and there are no relevant differences between the sporadic continuous manipulation in Case 1 and programming, then he may be able to undermine my argument. As I intend to show that Case 1 is not a counter-example in Chapter 6, I am not worried about this. I also apply and develop my argument against design cases (that is, the reverse generalisation strategy) to programming cases (and other similar cases) in Chapter 7.
Case 3: Plum [hereafter Plum3] 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 behavior 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 2001: 114)

Is Plum3 morally responsible for killing White? Plum3 has been conditioned by his parents and community to have the reasoning processes and desires that will eventually lead to him killing White. Conditioning is effectively a sort of programming, except that it is spread out over time. One difference between Cases 2 and 3 is that Plum3’s parents and community do not have knowledge of Plum3’s future environment. So it seems that although Plum3 is determined, he is not covertly controlled. This, it seems, is a relevant difference between Cases 2 and 3. However, it seems that Pereboom could easily modify this case to bring it into line with Case 2. Given that Plum3’s parents and community are (presumably) human agents, it is unfeasible for them to have complete knowledge of the future. But human agents could feasibly have local control over Plum3’s environment. Such control would ensure that Plum3 does not stray from his egoistic

---

116 Alternatively, we could leave Case 3 as it is and treat it like a causal determination case (like Case 4) and not as a control case. McKenna (forthcoming) does this. As I’ll argue shortly, the worry for the Four-Case Argument is that there is an obvious relevant difference between Case 4 and the other cases – namely Case 4 does not feature covert control (or manipulation, as figures in the debate tend to say). If we treat Case 3 more like Case 4, then the relevant difference obtains between Cases 1-2 and 3-4 rather than between Cases 1-3 and 4. I opt for the latter reading, though this case can be modified without affecting my argument. My reason for this reading is that Case 3, it seems to me, alludes to Walden Two-style cases, as discussed by Kane (1996: 65-69); according to Kane, such cases feature covert (nonconstraining) control, and I’m inclined to agree.
behaviour. Thus, Plum3 is controlled by his parents and community to the same extent that Plum2 is controlled by the neuroscientists. And if Plum2 is not morally responsible for killing White, then neither is Plum3. Again, it seems plausible that Plum3 satisfies the leading compatibilist conditions, and hence Case 3 is also a counter-example to those conditions.

Note that Case 3 is much more like a case of mere causal determination than Case 2 is. There are now no neuroscientists and there is no direct intervention into Plum3’s brain. Pereboom then removes all (apparent) covert control in his final case, a case of mere causal determination:

Case 4: Physicalist determinism is true, and Plum [hereafter Plum4] 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 organization 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 behavior 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 2001: 115)

Is Plum4 morally responsible for killing White? Initially a compatibilist might be sure that he is. However, it seems that there are no relevant differences between Plum3 and Plum4. The only difference between Cases 3 and 4 is that Case 3 makes vivid the causal origins of Plum3’s actions – namely, events beyond his control. By hypothesis, however, Plum4’s actions also originate in events beyond his control. Thus, there seem to be no relevant differences between Cases 1-3 and Case 4. Given that Plums 1-3 are not morally responsible, it seems to follow that Plum4 is also...
not morally responsible. And if Plum4 – a merely causally determined agent – is not morally responsible, then compatibilism must be false.

5.3 Manipulation arguments to the best explanation

In this section, I’m going to discuss some structural problems with the Four-Case Argument. As we’ll see, these are structural problems, not only for the Four-Case Argument, but for a particular sort of manipulation argument – namely one that requires a ‘best explanation’ premise to secure to falsity of compatibilism.

As Pereboom presents his argument (as described above, and at least to begin with – more on this shortly) it appears very much like any other manipulation argument – that is, it accords to the MAS, which I set out in §5.1. MAS, to repeat, goes like this:

M1. (Judgement) An individual covertly controlled in manner $X$ to perform action $A$ is not morally responsible for $A$-ing.

M2. (No difference) There are no differences relevant to moral responsibility between an individual who is covertly controlled in manner $X$ to $A$ and an individual who is causally determined to $A$.

Therefore,

M3. If causal determinism is true, then no one is morally responsible for anything they do. In other words, compatibilism is false.

As the Four-Case Argument includes multiple (alleged) counter-examples, it includes multiple ‘no difference’ premises. But, as we’ll see, this isn’t a significant difference. Below is a formal expression of the MAS version of the Four-Case Argument:
4CA-1. (Judgement) Plum1 is not morally responsible for killing White because he was covertly controlled via sporadic continuous manipulation to kill her.

4CA-2. (No difference) There are no differences relevant to moral responsibility between Plum1 and Plums 2-3.

Therefore,

4CA-3. Plums 2-3 are not morally responsible.

4CA-4. (No difference) There are no differences relevant to moral responsibility between Plums 1-3 and Plum4.

Therefore,

4CA-5. Plum4 (a causally determined individual) is not morally responsible.

Therefore,

4CA-6. If causal determinism is true, no one is morally responsible for anything they do. In other words, compatibilism is false.

Call this ‘4CA-MAS’. The crucial premise of the 4CA-MAS, as with any manipulation argument, is its first premise. If we do not judge that Plum1 is not morally responsible, then the argument does not get off the ground. (Of course, Pereboom claims that Case 2 could also get his argument off the ground, but that relies on Case 2 being able to independently pump the non-responsibility intuition – something which it seems far from clear that it can do – so let’s set aside this point aside until Chapter 7.) The Four-Case Argument, then, initially seems like any other manipulation argument – that is, an instance of MAS.

Things get murky at this point. Once Pereboom reaches Case 4 he doesn’t immediately conclude that compatibilism is false. Rather, he first considers an objection that he attributes to Lycan (1987/1995). Given that Plums 1-3 are covertly controlled and that Plum4 is merely causally determined, it seems that a relevant difference between Plums 1-3 and Plum4 is that the former were covertly controlled by other agents, whereas the latter is not. This seems to be a
relevant difference, and thus it seems the non-responsibility judgement does not generalise from Plums 1-3 to Plum 4 – that is, 4CA-4 is false. Call this the ‘no generalisation’ objection.\(^{117}\) According to Lycan, we need only posit a ‘no covert control’ condition on free will (more on this below). In response, however, Pereboom writes:

… the claim that this is a relevant difference is implausible. Imagine a further case
that is exactly the same as, say, Case 1 or Case 2, except that Plum’s states are
induced by a machine that is generated spontaneously, without intelligent design.
Would he then be morally responsible? The compatibilist might agree that this sort
of machine induction is responsibility-undermining as well, and then devise a
condition that stipulates that agents are not responsible for actions manipulated by
agents or machines. But this move is patently ad hoc. What explanation could there
be for the truth of this condition other than that it gets the compatibilist the result he
wants? (2001: 115-116)

Pereboom’s point is that the compatibilist lacks independent motivation for the claim that covert control by other agents is a relevant difference between Cases 1-3 and Case 4, and thus a ‘no covert control’ condition is untenable. If they claimed this was a relevant difference, they would have to make a similar claim about neural state induction by machines. He claims that we can replace the agent-manipulators (that is, the neuroscientists) in Cases 1 and 2 with machines

---

\(^{117}\) Kearns (2012) and King (2013) have both defended versions of the ‘no generalisation’ objection. While they both develop this objection by arguing that, instead of a covertly controlled individual not satisfying a condition on moral responsibility, manipulation arguments are question begging, they both make the same core point as Lycan (or at least the objection that Pereboom attributes to him); namely that we cannot generalise the non-responsibility judgement about a covertly controlled individual to a merely determined individual because covert control is essential in eliciting the non-responsibility judgement. Manipulation arguments beg the question, according to Kearns and King, by presuming the falsity of compatibilism to ensure the generalisation of the non-responsibility judgement. Pereboom’s appeal to the inference to the best explanation, then, is a way to avoid this objection. Todd (2013b) makes a similar move to Pereboom, though he is not as explicit about it as Pereboom is about what he doing.
without affecting our judgements that Plums 1 and 2 are not morally responsible (and presumably Pereboom thinks this can be done with Case 3 too).

Pereboom then asserts that:

The best explanation for the intuition that Plum is not morally responsible in the first three cases is that his action results from a deterministic causal process that traces back to factors beyond his control. Because Plum is also causally determined in this way in Case 4, we should conclude that here too Plum is not morally responsible for the same reason. More generally, if an action results from a deterministic causal process that traces back to factors beyond the agent’s control, then he is not morally responsible for it. (Pereboom 2001: 116; my emphasis)

This inference to the best explanation might seem a bit confusing. Pereboom writes as if he hasn’t yet shown that compatibilism is false and thus that he needs some further reason to establish that. But, by this point, hasn’t Pereboom, if his argument so far is sound, already shown that compatibilism is false? After all, it seems plausible that there are no relevant differences between Cases 1-3 and Case 4 (assuming that Pereboom is correct that covert control by other agents isn’t a relevant difference); so it should follow that Plum4 is not morally responsible, and hence that compatibilism is false. Adding an inference to the best explanation therefore seems unnecessary.

Although, ultimately, I think the inference to the best explanation is indeed unnecessary, it will be worth exploring why Pereboom takes it to be necessary. The answer lies with the objection that Pereboom discusses quickly before presenting his inference to the best explanation. Although Pereboom attributes this objection to Lycan (1987/1995), Lycan actually says very little on the matter. In response to manipulation cases of Taylor’s (1974) Lycan (1995: 117) says, ‘[w]hat we object to in these cases is precisely that the victim is the puppet of another
person – that his or her choices are coerced.’ Lycan then claims that we can add a negative condition – a ‘no covert control’ condition – to our analysis of free will that rules that agents who are the puppets of others are not free. Of course, as Lycan (1987/1995: 117) is well aware, such a condition seems ‘somewhat ad hoc’. So, as Pereboom notes above, it seems compatibilists cannot rely on such a condition.\(^{118}\)

While a ‘no covert control’ condition is unpromising, the ‘no generalisation’ objection can be pressed without claiming that there is such a condition on free will/moral responsibility.\(^{119}\) Instead of interpreting Lycan as proposing such a condition, we can interpret him as diagnosing the source (or primary cause) of the non-responsibility judgement about covertly controlled agents – viz. the fact they were covertly controlled. If it is the case that this fact is the source of the non-responsibility judgement, we have a prima facie reason to not generalise the non-responsibility judgement to Case 4; this effectively stops the Four-Case Argument in its tracks.

When Pereboom responds to this objection he seems worried that being the puppet of another person is a relevant difference that will stop his effort to straightforwardly generalise the non-responsibility judgement from Cases 1-3 to Case 4, and this is likely because he is aware that the fact Plums 1-3 were covertly controlled by other agents is important in eliciting the judgement that Plums 1-3 are not morally responsible. Why else would Pereboom include them in his cases if he didn’t think they were important in eliciting that judgement? Clearly, they are; but to avoid the ‘no generalisation’ objection it must be that covert control by other agents isn’t essential to eliciting the non-responsibility judgement. This, therefore, is the point of Pereboom’s claim that we can imagine cases just like Cases 1 and 2 except that the neuroscientists have been replaced with an intentionless machine that does all the same work as them. If we agree with Pereboom that Plum1* (an individual just like Plum1 except that his neural states were induced by an intentionless machine rather than a group of agents) is not morally responsible, then the

\(^{118}\) I discuss this condition again in §7.5.

\(^{119}\) This is how Kearns (2012) and King (2013) develop this objection. See n. 117 and §5.7.
fact that Plums 1-3 were covertly controlled by other agents isn’t in fact essential to eliciting the non-responsibility judgement.

The inference to the best explanation is deployed at this point. Its aim, it seems, is to ‘push through’ the generalisation of the non-responsibility judgement from Cases 1-3 to Case 4. Given that we (let’s assume) judge that Plums 1-3 are not morally responsible, it seems that we do so because they were covertly controlled. This, after all, is the common factor between the three of them. And given that it seems that Plum4 has not been covertly controlled, we can resist the generalisation of the non-responsibility judgement from Plums 1-3 to Plum4. Pereboom’s inference to the best explanation is, in effect, a way to overcome this resistance and to push through the generalisation. If Pereboom is correct that the best explanation for judgement that Plums 1-3 are not morally responsible is that their actions are the product of a deterministic causal process that traces back to factors beyond their control, then this entails that Plum4 is also not morally responsible as his actions are also the product of such a deterministic causal process.

It seems, then, that the Four-Case Argument is not an instance of MAS, but a different sort of manipulation argument, one which Mele (2008: 276) calls a ‘manipulation argument to the best explanation’. The logical form of a manipulation argument to best explanation, according to Mele, is the following (call it MABES):

M1. (Judgement) An individual covertly controlled in manner X to perform action A is not morally responsible for A-ing.

M2*. (Best explanation) The best explanation for the non-responsibility judgement about a covertly controlled individual is that she has been causally determined by events external to her.

Therefore,

M3. If causal determinism is true, then no one is morally responsibility for anything they do. In other words, compatibilism is false.
MABES is just like MAS except that it replaces the ‘no difference’ premise with a best explanation premise. The Four-Case Argument interpreted as instance of MABES looks like this:

4CA-1. (Judgement) Plum1 is not morally responsible for killing White because he was covertly controlled via sporadic continuous manipulation to kill her.

4CA-2. (No difference) There are no differences relevant to moral responsibility between Plum1 and Plums 2-3.

Therefore,

4CA-3. Plums 2-3 are not morally responsible.

4CA-IBE. (Best explanation) The best explanation for the non-responsibility judgement about Plums 1-3 is that they have each been causally determined by events external to them.

Therefore,

4CA-5. Plum4 (a causally determined individual) is not morally responsible.

Therefore,

4CA-6. If causal determinism is true, then no one is morally responsible for anything they do. In other words, compatibilism is false.

Call this argument ‘4CA-MABES’. Note that this still includes a ‘no difference’ premise, 4CA-2. Pereboom, after all, argues that Plums 2 and 3 are not morally responsible on the basis that Plum1 is not and that there are no relevant differences between Plum1 and Plums 2 and 3. However, the ‘no difference’ premise is not being deployed in the move to show that Plum4 is not morally responsible. That move is secured instead by the best-explanation premise.

But there is a problem with this interpretation of the Four-Case Argument. Pereboom doesn’t actually replace the ‘no difference’ premise 4CA-4 with a best explanation premise like
4CA-IBE. Rather, he introduces the best explanation premise to *support* the no-difference premise. Thus the Four-Case Argument actually runs like this:

4CA-1. (Judgement) Plum1 is not morally responsible for killing White because he was covertly controlled via sporadic continuous manipulation to kill her.

4CA-2. (No difference) There are no differences relevant to moral responsibility between Plum1 and Plums 2-3.

Therefore,

4CA-3. Plums 2-3 are not morally responsible.

4CA-IBE. (Best explanation) The best explanation for the non-responsibility judgement about Plums 1-3 is that they each been causally determined by events external to them.

Therefore,

4CA-4. (No difference) There are no differences relevant to moral responsibility between Plums 1-3 and Plum4.

Therefore,

4CA-5. Plum4 (a causally determined individual) is not morally responsible.

Therefore,

4CA-6. If causal determinism is true, then no is morally responsible for anything they do. In other words, compatibilism is false.

Call this argument ‘4CA-MAS-IBE’. 4CA-MAS-IBE is an instance of MAS. The only difference is that the 4CA-MAS-IBE is much more explicit in what support it has for its crucial ‘no difference’ premise – viz. the best explanation premise (4CA-IBE).

What about MABES/4CA-MABES, then? This seems to remain a distinct kind of manipulation argument. However, among other things, I am going to argue that 4CA-IBE is a
redundant premise in 4CA-MAS-IBE because it is not required to show that 4CA-4 is true – that is, it can be established that there are no relevant differences between Plums 1-3 and Plum4 without making an inference to the best explanation of our judgements. Hence, 4CA-MAS-IBE collapses into 4CA-MAS. I will show that an incompatibilism-conducive inference to the best explanation is only possible if there are no relevant differences between a manipulated and determined individual. Thus, the ‘no difference’ premise actually supports the ‘best explanation’ premise, and not vice versa as Pereboom claims. But if a ‘no difference’ premise is required to support a ‘best-explanation’ premise, then the ‘best explanation’ premise is superfluous; after all, if the ‘no difference’ premise (and of course the judgement premise) is true, then it follows that compatibilism is false. MABES/4CA-MABES merely suppresses the ‘no difference’ premise, and so isn’t a different type of manipulation argument to MAS/4CA-MAS-IBE.\textsuperscript{120}

5.4 Covert control is the best explanation

There are three problems with 4CA-MAS-IBE. First, it is not clear that Pereboom’s best explanation really is the best explanation. Second, to secure the conclusion that compatibilism is false, the inference to the best explanation relies on us \textit{not} having judgements about \textit{why} Plums 1-3 are not morally responsible. Third, the inference to the best explanation is redundant; the ‘no difference’ premise supports the ‘best-explanation’ premise (and not vice versa), and the ‘no difference’ premise (in conjunction with the judgement premise) is sufficient to show that compatibilism is false. In this section, I focus on the first problem. I discuss the second and third problems in §5.5 and §5.6 respectively.

The point of the inference to the best explanation is to show that there are in fact no relevant differences between Plums 1-3 and Plum4. Given that the (alleged) best explanation for the intuition that Plums 1-3 are not morally responsible is that they are causally determined by

\textsuperscript{120} One manipulation argument which seems to necessarily include a best explanation premise is Todd’s (2012) ‘moral standing’ manipulation argument. I discuss this argument in §7.4.
events external to them, it seems to follow that Plum4, a causally determined individual, is also not morally responsible. In effect, Pereboom is able to ‘push through’ the non-responsibility judgement about Plums 1-3 so that it applies to Plum4 too.

As we saw in the previous section, the inference to the best explanation premise (4CA-IBE) has its roots in Pereboom’s response to the ‘no generalisation’ objection. Pereboom responds to this objection by positing further cases, which feature machines rather than agent-manipulators. Pereboom’s own analysis of the prima facie cause of the non-responsibility judgement is not that Plums 1-3 have been causally determined; rather what is prima facie driving that judgement, according to Pereboom, is ‘causal determination by other agents’ (Pereboom 2014: 79). Supposing we do judge that Plums 1-3 are not morally responsible because they have been causally determined by other agents, the point of Pereboom’s additional cases is to show us ‘that causal determination by other agents was not essential to what was driving the intuition of non-responsibility’ (Pereboom 2014: 79).

Pereboom’s view seems to be that inference to the best explanation actually underpins the claim there are no relevant differences between Plums 1-3 and Plum4, and that’s why he thinks the ‘real structure’ of manipulation arguments includes a best-explanation premise (Pereboom 2014: 79-80, n.3). On Pereboom’s view, then, the reason that there are no relevant differences between Cases 1-3 and Case 4 is that the best explanation of the non-responsibility of Plums 1-3 shows that there are no relevant differences between Plums 1-3 and Plum4. To establish this, though, Pereboom must appeal to further cases featuring machines inducing the mental states of Plum1* and Plum2*. By appealing to further cases, Pereboom claims to show that covert control by other agents is not what is driving the non-responsibility judgement about Plums 1-3.

But if we are allowed to expand the pool of cases to include agents causally determined by intentionless forces, we are surely allowed to expand our pool of cases to include indeterministic agents who have been covertly controlled as well. What happens if we do this instead? Well, as Mele (2005) points out, it seems intuitively plausible that indeterministic agents
are not morally responsible in cases analogous to Cases 1-3 (Cases 1a-3a). In these cases, Plums 1a-3a are covertly controlled just like Plums 1-3 to kill White. The only difference is that while Plums 1-3 are guaranteed to kill White, there is a small chance that Plums 1a-3a will not kill White because there is a small chance they will have a mental breakdown. Given this, the best explanation for our intuitions about covertly controlled individuals cannot be they have been causally determined by events beyond their control because that would not explain why we judge that indeterministic individuals are also not morally responsible when they are covertly controlled. After all, by definition, indeterministic individuals are not causally determined by events beyond their control. In other words, if we expand our range of comparison cases to include Mele’s indeterministically-controlled individuals as well, Pereboom’s best explanation is no longer in the running, so we need to look elsewhere for our best explanation for the non-responsibility of all of these agents.

Mele himself concludes that the fact all these agents have been manipulated is the best explanation for the non-responsibility intuition. This, I think, is a little imprecise. To say something has been manipulated is rather ambiguous. If Mele’s best explanation for the non-responsibility intuition is correct, then it ought to follow that manipulation is necessarily responsibility-undermining. But this is false. Neuroscientists could manipulate a person by removing some of her memories. Would this person lack moral responsibility for her subsequent actions? Clearly not. Hence, we must specify what exactly about manipulation is responsibility-undermining. What is worrisome about manipulation is, as Lycan notes, the puppet factor – that is, the possibility that an agent may be covertly controlled to perform her actions. While an individual who has some memories removed has been manipulated, she has not been covertly controlled; there is nothing about memory manipulation that provides the manipulator with control over the manipulatee’s actions. Thus, the best explanation for the non-responsibility judgement about all of Plums 1-3 (the deterministic, agent-manipulated Plums), Plums 1*-3* (the
deterministic, machine-induced Plums) and Plums 1a-3a (the indeterministic, agent-manipulated Plums) would seem to be that they have been covertly controlled.

A critic might object to my analysis because it implies that machines can covertly control human beings. Such a critic may also note that Pereboom stipulates that those machines are intentionless, and then infer that they cannot control anything. But intentions and other mental states are not necessary for control. Intentions and other mental states are certainly necessary for a particular type of control – namely, the control required of an agent for their own action and free action. But intentionless machines control all sorts of things: traffic lights, environmental controls, production lines, etc. Of course, these machines are all designed for controlling by agents, so this analogy might be problematic. However, we can imagine an intentionless machine that forms spontaneously in space without any intelligent design. If a machine that controlled a production line were replaced (somehow) with a qualitatively identical duplicate that formed spontaneously in space without any intelligent design, then it would still be controlling the production line, but without any agents in its causal history. Pereboom, in fact, asks us to imagine such an intentionless machine replacing the neuroscientists in his Case 1*. It seems that if Plum1* is not morally responsible for killing White it is because the machine is covertly controlling him by virtue of inducing the neural states that underlie the reasoning process that leads to him killing White.

Hence, Pereboom hasn’t provided the best explanation for the cause of the non-responsibility judgement. Although he concedes that initially the best explanation for the non-responsibility judgement might be that Plums 1-3 have been causally determined by other agents, Pereboom aims to undercut that view by proposing Cases 1*-3*, which feature no agent-manipulators. So, it seems that causal determination by other agents is not the best explanation for the non-responsibility judgement, but rather mere causal determination is. However, once we consider cases of indeterministic covert control, it becomes clear that causal determination cannot be the best explanation. Although Mele (2005) claims that the fact all the Plums have been
manipulated is the best explanation, I have argued that the best explanation for the non-responsibility judgement is that they have been covertly controlled.

5.5 Why is Plum not morally responsible?

It would appear that, according to Pereboom, the inference to the best explanation is pivotal in pushing through the generalisation of the non-responsibility judgement from Cases 1-3 to Case 4. In this section, I diagnose why Pereboom has incorrectly identified the best explanation as causal determination by external events, and I then consider a (perhaps implausible) way for incompatibilists to avoid the ‘no generalisation’ objection.

The reason that Pereboom has not identified the best explanation is that he ignores the fact that when we judge that Plums 1-3 are not morally responsible, we also judge why they are not morally responsible. So, it’s not the case that we only judge that:

(1) Plum1 is not morally responsible.

We also have a judgement about why Plum1 is not morally responsible. I contend that we also judge that:

(2) Plum1 is not morally responsible because he has been covertly controlled.\(^{121}\)

When, as I argued in §5.2, it seems that Plum2 or Plum3 have not been covertly controlled because they have only been programmed or conditioned (that is, programming or conditioning have not been combined with knowledge or local control of Plum2’s and Plum3’s respective environments), this seems like a relevant difference between them and Plum1. Given that there

\(^{121}\) Another way to put it is that we first judge that Plum1 has been covertly controlled, and then we infer that Plum1 is not morally responsible from that judgement.
seems to be a relevant difference between Plum1 and Plums 2 and 4, the non-responsibility judgement does not generalise from Plum1 to Plums 2 and 3, and so those cases must be modified accordingly. Thus, in response to Cases 1-3 (with Cases 2 and 3 suitably modified), we judge:

(1*) Plums 1-3 are not morally responsible.

And we also judge that:

(2*) Plums 1-3 are not morally responsible because they have been covertly controlled.

(1*) is the non-responsibility judgement. (2*), on the other hand, is an explanatory or diagnostic judgement; it is a judgement about why Plums 1-3 are not morally responsible, and it conforms to the best explanation of the non-responsibility judgement. Thus, it seems plausible that the best explanation of the non-responsibility judgement informs us of what we judge makes Plums 1-3 not morally responsible. Since covert control is the most salient feature of a whole range of cases, it seems that covert control is what we judge makes Plums 1-3 not morally responsible. Of course, given that Plum4 (it seems) has not been covertly controlled, there seems to be a relevant difference between Plums 1-3 and Plum4 – viz. Plums 1-3 have been covertly controlled and Plum4 has not.

But perhaps Pereboom can claim, as Todd (2013b) recently has, that manipulation (and presumably, covert control) is irrelevant to what actually makes a manipulated/covertly controlled agent non-responsible, even though it might be what we intuitively judge to make a covertly controlled individual non-responsible. Todd writes:
... the proponent of [a manipulation] argument should admit that the manipulation does no work in making the agent unfree [or non-responsible]. Rather, the proponent of the argument contends – and clearly must contend – that the manipulation is irrelevant as concerns what makes the agent unfree. She instead says that the manipulation can help us see that something does make the agent unfree. In other words, she first presents the scenario (say) to an agnostic, and asks whether the agnostic thinks that the agent is free (or responsible) in that scenario. And suppose the agnostic says ‘no’. She then points out that whatever would make the agent unfree in that scenario would also make the agent unfree in a qualitatively identical scenario, except in which blind natural causes have taken the place of an intentional agent. (Todd 2013b: 202; my emphasis)122

The idea is that after an agnostic (or neutral inquirer) has been presented with a control case they judge that the featured individual is not morally responsible. So, suppose an incompatibilist presents Cases 1-3 to an agnostic, and the agnostic then judges that Plums 1-3 are not morally responsible. Todd claims that the incompatibilist can then tell the agnostic that the covert control was not, in fact, relevant to why Plums 1-3 are not morally responsible. The incompatibilist is then able to provide an alternative explanation which is conducive to incompatibilism (i.e. Plums 1-3 are not morally responsible because they have been causally determined by events beyond their control). In effect, it seems that Todd is trying to push the generalisation from a control case to a determination case (such as from Cases 1-3 to Case 4) by offering an explanation akin to Pereboom’s inference to the best explanation.

The trouble for Todd is that he doesn’t have an argument for why causal determination makes covertly controlled individuals, such as Plums 1-3, not morally responsible. Unlike

---

122 Todd is actually talking about the Designed Ernie case. I take it that he thinks these points apply to all manipulated/covertly controlled individuals.
Pereboom, Todd doesn’t present any further cases in which individuals are covertly controlled (or manipulated) by intentionless machines or some other blind force. Todd (2013b: 202) does note that ‘whatever would make the agent unfree in that scenario would also make the agent unfree in a qualitatively identical scenario, except in which blind natural causes have taken the place of an intentional agent.’ But the qualitatively identical scenario he has in mind, it seems, is a case of mere causal determination, such as Case 4. Hence, we have no reason to believe that (a) causal determination makes Plums 1-3 not morally responsible, or that (b) manipulation/covert control is irrelevant to what makes Plums 1-3 not morally responsible.

The only argument available for (a) and (b), however, is further cases, such as Cases 1*-3* – that is, cases without agent-manipulators. Such cases provide a reason to discount the claim that covert control by other agents is essential to the non-responsibility judgement, but they do not discount the claim that covert control is essential to the non-responsibility judgement. It seems that any case the incompatibilist presents will have to include covert control of some kind if she wishes it to elicit the non-responsibility judgement. Pereboom’s Case 2 is a good example of this. As I argued in §5.2, prior to it being made clear that Plum2 has been covertly controlled (via programming in conjunction with knowledge or local control of Plum2’s environment) by the neuroscientists, it does not seem implausible to say that Plum2 is morally responsible. Thus, in programming cases, additional factors, such as knowledge of the future or local control of an individual’s environment, in conjunction with programming are required to make it intuitively plausible that an agent is not morally responsible. The reason it is intuitively plausible that Plum2 is not morally responsible is because he has been covertly controlled. This point generalises to Case 3.\textsuperscript{123}

Todd might try to present some further cases apparently without covert control – perhaps one featuring ‘blind natural causes’ rather than intentionless machines. But this would

\textsuperscript{123} I discuss this point more in §7.2.
not help because blind natural causes can still control things. For example, the wind can control the movement of trees; the sun controls the growth of plants; gravitational forces control planets, insulin controls our blood sugar levels, nerves control our bodily movements, and so on. The wind, the sun, gravity, insulin, and nerves are not agents yet it makes perfect sense to say that they control things. And we can even imagine someone using, ‘I didn’t mean to push Jerry off the bridge; the wind made me do it!’ as excuse for pushing Jerry from a bridge. Of course, we might worry that these natural forces do not covertly control things. Perhaps these natural forces do not, but we might imagine natural forces that do. Mele (2005: 77) imagines another series of cases analogous to Cases 1-3 except that a ‘spontaneously generated electromagnetic field’ is doing the work the agent-manipulators did in Cases 1-3 (call these Cases 1b-3b). Now, just as the machine covertly controls Plums 1*-3*, it seems that the electromagnetic field covertly controls Plums 1b-3b. So, blind forces can covertly control individuals, and so there are still no grounds for the claim that causal determination is what makes Plums 1-3 not morally responsible.

Incompatibilists have yet to devise a scenario that elicits the judgement that an agent is not morally responsible, and yet where the agent is not covertly controlled by something. Moreover, it seems unlikely that they ever could. Any case that stops short of covert control will not get a manipulation argument off the ground because, as I’ve argued, such a case will not elicit the non-responsibility judgement. And it seems that incompatibilists are (at least) implicitly aware of this fact, because they only ever use cases featuring covert control to get their manipulation arguments off the ground. If Pereboom thought that his machine-induction cases (Cases 1*-3*) were just as compelling as his Cases 1-3, then why didn’t he just use those? The fact he didn’t is telling. The fact incompatibilists have never done this is even more telling.

So, covert control is essential to eliciting the non-responsibility judgement; this impacts on both the ‘no difference’ premise and the ‘best explanation’ premise. The former premise requires that it is plausible that there are no relevant differences between a covertly controlled individual and a merely determined individual. But if covert control is essential in eliciting the
non-responsibility judgement, then there is (it seems) a relevant difference between a covertly controlled individual and a merely determined individual – namely, a covertly controlled individual has been covertly controlled while a merely determined individual has not. Similarly, if covert control is essential in eliciting the non-responsibility judgement, the best explanation of a covertly controlled individual’s non-responsibility is that she has been covertly controlled. Hence, both ‘straight’ manipulation arguments and manipulation arguments to the best explanation fail because they are unable to overcome the ‘no generalisation’ objection.

But there is a way for incompatibilist to respond to the no generalisation objection without taking the Todd/Pereboom route. Incompatibilists could argue that there are no relevant differences between Cases 1-3 and Case 4 because while Plums 1-3 are covertly controlled by other agents, Plum4 is *covertly controlled by events external to him*. Perhaps this generalisation could be pushed through with an inference to the best explanation: namely, the best explanation for the non-responsibility of Plums 1-3 is that they have been covertly controlled. Given that Plum4 is also covertly controlled by other things (such as a combination of his upbringing, his genetics, people around him, the weather, the big bang – in other words, events external to him), he is not morally responsible. This control is covert insofar as Plum4 is not aware of the causal role these events have over his subsequent actions. Hence, the ‘no generalisation’ objection is unsuccessful as long as incompatibilists construe the conclusion of manipulation arguments in the following way:

M3*: If causal determinism is true, no one is morally responsible for anything they do *because* all causally determined individuals are covertly controlled by events external to them.

Certain incompatibilists won’t like M3* as a conclusion. The upshot is that the ultimate explanation of why deterministic individuals lack moral responsibility is not because they have
been causally determined by events beyond their control, but rather that they have been *covertly controlled*. Although manipulation arguments (if sound) would show that compatibilism is false, they would not show that compatibilism is false because determinism precludes responsibility; rather, they would show that compatibilism is false because being in the world with deterministic laws results in individuals being covertly controlled by events external to them. The reason that some incompatibilists won’t like it is that many of them (including Pereboom and Todd) are committed to (at least) the metaphysical possibility of free will and moral responsibility. If it turns out that compatibilism is false because being in the world with deterministic laws results in individuals being covertly controlled by external events, then it seems that being in a world with *indeterministic* (that is, probabilistic) laws would also entail that individuals are covertly controlled by external events. This suggests, as Mele argues, that the responsibility-undermining feature of control cases is not deterministic laws *per se* but that responsibility requires that we *not* be covertly controlled.

### 5.6 The redundancy of the inference to the best explanation

In this section, I argue that the inference to best explanation is not required to support the conclusion that compatibilism is false. The inference is supposed to show that there are not actually any relevant differences between Cases 1-3 and Case 4. However, Pereboom deployes the inference to the best explanation *after* arguing that an alleged relevant difference between Cases 1-3 and Case 4 is not, in fact, a relevant difference. If there aren’t actually any relevant differences between Cases 1-3 and Case 4, then the inference to the best explanation isn’t required, because if there are no relevant differences between Cases 1-3 and Case 4, then it follows straightforwardly that Plum4 is not morally responsible. The only apparent relevant difference between Cases 1-3 and Case 4 is that Plums 1-3 have been covertly controlled by other agents, while Plum4 has not. But Pereboom’s further cases (Case 1*-3*, which replace the
neuroscientists with machines) show that covert control by other agents is not a relevant difference; so there is no need for the inference to the best explanation.

In recent work, though, Pereboom says the following in support of his view that manipulation arguments require a ‘best explanation’ premise:

Al Mele (2006) argues that a manipulation argument against compatibilism need not be cast as an argument to the best explanation. I doubt that this is so. True, the argument can be represented without a best-explanation premise, but such a representation will not reveal its real structure. By analogy, the teleological argument for God’s existence can be represented as a deductive argument, but its real structure is an argument to the best explanation for biological order in the universe. The fact that the real structure of a manipulation argument against compatibilism is an argument to the best explanation becomes clear when one considers compatibilist objections to it – that, for example, the non-responsibility intuition can be accounted for by manipulation of a certain sort and not by causal determination [the ‘no generalisation’ objection]. (Pereboom 2014: 79-80, n.3)

This is Pereboom’s only explicit support for the claim that the inference to the best explanation is essential to any manipulation argument, and therefore that the ‘real’ structure of manipulation arguments must include a best explanation premise. I take that what Pereboom means is that the ‘best explanation’ premise supports the ‘no difference’ premise. So, in order to show that there are no relevant differences between Plums 1-3 and Plum4, it must be the case that the best explanation of the intuition that Plums 1-3 are not morally responsible is that they have been causally determined by events external to them. Pereboom then claims that manipulation arguments are similar to teleological arguments for the existence of God with respect to the apparent fact that both require a ‘best explanation’ premise.
But this analogy actually works against Pereboom. Contrary to what Pereboom implicitly claims, the best explanation premise of a teleological argument is only true if the analogy between examples of human design and (alleged) divine design holds. Deductive teleological arguments go (roughly) like this. Here’s something designed by people (e.g. a watch); it’s really complex (etc.). Here’s the world; it’s really complex (etc.) too. Like effects must have like causes; so if a watch has a designer then the world must have a designer. Hence, the world has a designer (or designers). Central to this argument, then, is the analogy between human and (alleged) divine products. Because of the similarities between these products, it apparently follows that if one is designed then the other one must be too.

Best-explanation teleological arguments, on the other hand, apparently do not appeal to the analogy between human and divine products. Instead, they replace the analogy premise with a best-explanation premise. Thus, best-explanation arguments go (roughly) like this. Here’s the world; it’s really complex (etc.). The best explanation for the complexity of the world is that it has been designed. Hence, a designer exists. Supposing that we agree that the best explanation for complexity (or biological order, as Pereboom puts it) of the world is that it has been designed, why does it seem that the existence of a designer is the best explanation for such complexity? To say that something is designed, we must implicitly be comparing it to some examples of design we have. If we didn’t, we would have no reason to infer that the existence of a designer is the best explanation. But the only uncontroversial experience we have of design is of human design. Why else would something’s complexity make us think it had been designed if we were not thinking of other complex things that we know have been designed? To judge that the best explanation of X is design therefore requires that we think X is like human design. This inference to the best explanation, then, requires accepting the analogy between human and

---

124 I intend to develop this argument into a paper at some point. In short, contemporary proponents argue that Hume’s criticisms do not undermine a best explanation design argument (see, for example, Pruss and Gale 2005: 129). My argument, while admittedly sketchy and in need of further development, shows that Hume’s criticisms are just as effective against best explanation design arguments as they are against analogical design arguments.
divine products. Hence, the central premise of deductive teleological argument (the analogy premise) is built into best explanation teleological arguments. (And that’s bad news for defenders of best explanation teleological arguments because these arguments consequently inherit all the problems with deductive teleological arguments.) If the ‘best-explanation’ premise is true, then the analogy between human products and (alleged) divine products holds. But if the analogy between human products and (alleged) divine product holds, then we can show God exists without including a ‘best-explanation’ premise in our design argument. Hence the ‘best-explanation’ premise is not required. Best-explanation teleological arguments therefore collapse into deductive teleological arguments.

A similar point applies to manipulation arguments. Consider a straight manipulation argument. First, it must be argued that a covertly controlled individual is not morally responsible. Second, it is argued that there are no relevant differences between the covertly controlled individual and a merely causally determined analogue of that covertly controlled individual. From these two premises, incompatibilists claim that it follows that compatibilism is false. Central to this argument, it seems, is the ‘no difference’ premise. Manipulation arguments to the best explanation, however, apparently replace the ‘no difference’ premise with a ‘best-explanation’ premise: it is argued instead that the best explanation for the judgement that the covertly controlled individual is not morally responsible is that she has been causally determined by events external to her. From this premise and the first premise of the straight manipulation argument, incompatibilists claim it follows that compatibilism is false. But why should we think that causal determination is the best explanation of the non-responsibility judgement? Presumably we don’t start by judging that causally determined individuals are not morally responsible; otherwise, there would be no point in presenting the argument. So, we must either have an opposing judgement or no judgement at all. So there seems to be no reason to infer that causal determination by external events is the best explanation of the non-responsibility judgement – unless we’re already convinced that there are no relevant differences between a
covertly controlled individual and a merely causally determined individual. So just as the central premise of the deductive teleological argument is built into a best explanation teleological argument, the central premise of a straight manipulation argument is built into a manipulation argument to the best explanation. Since the only support for the ‘best-explanation’ premise is the truth of the ‘no difference’ premise, the ‘best-explanation’ premise cannot be used to support the ‘no difference’ premise. Hence, manipulation arguments to the best explanation collapse into straight manipulation arguments.

So, Pereboom’s analogy between manipulation arguments and teleological arguments in fact works against him. The only support he offers for his view that the ‘real’ structure of manipulation requires a best explanation premise is unpersuasive. Indeed, the comparison between teleological and manipulation arguments shows that ‘best explanation’ premise is redundant. I’ve argued that we can’t establish the truth of the ‘best explanation’ premise until we have established the truth of the ‘no difference’ premise. And if we accept that there are no relevant differences between Plums 1-3 and Plum4, then it follows that compatibilism is false (because Plum4 is not morally responsible) without there being any need to make an inference to best explanation.

5.7 The control argument

In §5.5, I considered one way for incompatibilists to avoid the ‘no generalisation’ objection. I proposed that incompatibilists could claim that merely determined individuals are also covertly controlled, and so there is no problem with generalising the non-responsibility judgement about a covertly controlled individual to a merely determined individual. In this section, I wish to outline a different way for incompatibilists to resist the ‘no generalisation’ objection. The upshot is that it will then be clear what compatibilists need to do to overcome the threat posed by control cases.
The ‘no generalisation’ objection shows that instances of MAS fail to show that compatibilism is false by arguing that M2 is false because there is a relevant difference between an individual covertly controlled in a particular way and a merely determined individual. Indeed, while this objection can be construed such that it posits a ‘no covert control’ condition on moral responsibility (which is how Lycan construes it), it can in fact be presented without claiming that the merely determined individual satisfies some condition that the covertly controlled individual fails to satisfy. Kearns (2012) and King (2013) both develop the objection in this way (and this is how I developed the objection in §5.4). They argue that we cannot generalise our judgement about a covertly controlled individual to a merely determined individual. The judgement about the former depends on there being covert control (or manipulation, as they both put it) in a given scenario. Since covert control is missing from a mere determination scenario, they claim that we cannot generalise the non-responsibility judgement to such cases. Indeed, Kearns and King both claim that manipulation arguments beg the question against the compatibilist: according to them both, M2 only seems true (for a particular instance of MAS) if we’ve already accepted that compatibilism is false, and such reasoning is fallacious.

Let’s assume for the sake of argument that the ‘no generalisation’ objection succeeds. Have Kearns and King saved compatibilism? No. Even if manipulation arguments (that is, instances of MAS) do not show that compatibilism is false, there is still the problem of the apparent counter-examples to the leading compatibilist conditions on moral responsibility. Remember Pereboom’s Cases 1-3 seem to be prima facie counter-examples to the leading compatibilist conditions on moral responsibility. They therefore constitute objections to compatibilism, in its current form, in their own right. Defenders of compatibilism must still, therefore, overcome these objections in order to maintain their position. Of course, manipulation arguments attempt to show that the apparent responsibility-undermining features of control cases is a feature of any determination case. In other words, there’s no point in coming up with new conditions on moral responsibility – according to manipulation arguments –
because the problem is with causal determinism, and not the leading compatibilist conditions. Defenders of the ‘no generalisation’ objection seem to think that compatibilists can rest easy because they have shown that our judgements do not suggest that causal determinism is the responsibility-undermining feature of control cases, and therefore do not support an argument against compatibilism. This, in effect, shows that our judgements do not rule out some as yet unexpressed compatibilist conditions on moral responsibility. But given that there are counter-examples to the leading compatibilists conditions on moral responsibility, incompatibilists have shown that these leading conditions are insufficient for moral responsibility. Indeed, it seems that incompatibilists have (apparent) counter-examples to all plausible current compatibilist conditions on moral responsibility, which includes my structural-narrative view, such as Pereboom’s Cases 1-3. Given this, compatibilists have a problem.

But is the problem significant enough to undermine compatibilism? It seems so, or so I shall now argue. Let’s think again about the historicist’s (unsuccessful) argument against structuralism. Historicists, recall, used brainwashing cases, such as Brainwashed Beth, to argue that structuralism is false. (For the next few paragraphs, I’ll talk as if the Brainwashed Beth case is still viable, but bear in mind that I argued in Chapter 4 that this case is not in fact a counter-example to my structural-narrative view.) Brainwashed Beth features an agent, Beth, who is brainwashed to have the system of values of a notorious serial killer, Chuck. Despite satisfying the structuralist conditions on moral responsibility, it seems that Beth is not morally responsible. Thus, it seems that Brainwashed Beth is a counter-example to the structuralist conditions, and therefore that structuralism is false. Of course, if the historicist believes the Brainwashed Beth shows that structuralism is false, she clearly must think that this case supports some argument against structuralism. This argument seems to go like this:

H1. Brainwashed Beth is a counter-example to the structuralist conditions on moral responsibility (i.e. synchronic ownership).
Therefore,

H2. Structuralism is false.

This argument seems valid and, if it is sound, it shows that structuralism is false. Notice that this argument contains nothing like the ‘no difference’ premise. Thus, positions can be shown to be false without a ‘no difference’ premise. This sets the stage for an argument against compatibilism that eschews a ‘no difference’ premise, thereby avoiding the ‘no generalisation’ objection. The argument goes like this:

C1. [Insert case here] is a counter-example to all plausible current compatibilist conditions on moral responsibility.

Therefore,

C2. Compatibilism is false.

Call this ‘the control argument’, as it is supported solely by a control case. A worry with the control argument (or any of its instances) is that it is invalid; it might seem that it does not follow that compatibilism is false just because there is a counter-example to all plausible current compatibilist conditions on moral responsibility. Mele, for instance, writes:

125 Notice that this argument does not yet motivate historicism. In attempt to establish that, historicists compare Beth with a non-brainwashed, though causally determined, individual, Chuck. The historicist assumes that Chuck is morally responsible (an assumption that is permissible in a debate between compatibilists). Granting this assumption, it seems that the only relevant difference between Beth and Chuck is a historical one; hence there must be a historical condition on moral responsibility that Chuck, but not Beth, has satisfied. This move, I contend, is an implicit inference to the best explanation. By comparing Beth with Chuck, the historicist seeks to establish why Beth is not morally responsible – that is, she seeks to explain the (apparent) fact that Beth is not morally responsible. Obviously, an incompatibilist would claim that it is because Beth has been causally determined. However, the historicist rules out this possible explanation of Beth’s non-responsibility since she has assumed that moral responsibility is compatible with determinism. This therefore leaves only the claim that Beth has not satisfied a historical condition on moral responsibility as the best explanation of her non-responsibility, thereby motivating historicism. The overall argument for historicism, then, includes an inference to the best explanation. Notice that manipulation arguments conclude that compatibilism is false, and so do not automatically support a positive form of incompatibilism (e.g. source or leeway incompatibilism). It seems that only an inference to the best explanation could support positive positions; hence inferences to the best explanation do have a role to play in manipulation arguments after all.
Of course, even if [a control case] is a counterexample to ['all alleged sets of conceptually sufficient conditions for free and morally responsible action ever proposed by compatibilists'], it does not follow that incompatibilism is true. Perhaps [a control case] is not a counterexample to some superior candidate for being a set of a conceptually sufficient conditions for free action and moral responsibility that is consistent with the truth of causal determinism and that has not yet been proposed by compatibilists. (Mele 2008: 264)

Because a control counter-example to all plausible current compatibilist conditions leaves open the possibility of some as yet unexpressed compatibilist conditions, such a control counter-example apparently cannot show that compatibilism is false. An initial problem for Mele is the following. If he’s correct, then he has no argument against structuralism. A structuralist could claim – just as Mele does with respect to control counter-examples to compatibilism – that a control counter-example to structuralism is only a counter-example to the current structuralist conditions; it does not rule out some as-yet-unexpressed structuralist conditions, and so the falsity of structuralism has not been established.

Such a defence of structuralism would be cheap, though – for surely the burden is on the structuralist to explain why apparent counter-examples to the structuralist conditions are not, in fact, counter-examples. This might involve positing further conditions (as I did in Chapter 3 in response to hypnosis cases) or might involve arguing that covertly controlled individuals are morally responsible (as I did in Chapter 4 in response to brainwashing cases). Unless and until the structuralist shows why the apparent counter-examples are not really counter-examples, it seems acceptable to conclude that structuralism is false. (Of course, as noted, I’ve argued that structuralists can overcome these problems; so structuralism has not been shown to be false.)
Likewise, such a defence of compatibilism in response to a control/manipulation argument would be cheap. If there is a counter-example to all plausible current compatibilist conditions on moral responsibility, then it seems that the burden is on the compatibilist to respond to such cases. Unless and until compatibilists respond to such cases, it seems acceptable to conclude that compatibilism is false.

Indeed, there is something strange about Mele’s requirement that, in order to show that compatibilism is false, a control case must be a counter-example to all possible compatibilist conditions on moral responsibility. If this were a general rule about counter-examples to positions, then it seems no counter-examples ever undermine philosophical positions. After all, it seems that we will never be in an epistemic position such that we can rule out some further possible conditions on something; we’re not omniscient! This is an implausible result, and it highlights that the condition that Mele places upon counter-examples is far too strong. That a control case is a counter-example to all plausible current compatibilist conditions on moral responsibility constitutes a sufficient reason to reject compatibilism. This, of course, does not rule out as-yet-unexpressed compatibilist conditions on moral responsibility. But the mere possibility that there are such conditions, just by itself, shouldn’t be taken to undermine the counter-example.

It should be clear, then, that as long as there are control counter-examples to the compatibilist conditions on moral responsibility, that compatibilism is in danger. The compatibilist’s burden is to show why these cases are not, in fact, counter-examples.

5.8 Conclusion

In this chapter I have sought to clarify the mechanics of manipulation arguments through an in depth examination of the Four-Case Argument. First, I argued the ‘no generalisation’ objection undermines manipulation arguments unless incompatibilists modify their conclusion, though I noted that incompatibilists may dislike this conclusion because it might ultimately undermine
their positive accounts of moral responsibility. Second, I argued ‘best explanation’ premises are not required to show that compatibilism is false. Thus, I showed that the Four-Case Argument does not need such a premise, and so 4CA-MAS-IBE collapses into 4CA-MAS. Third, I argued that control cases support another argument against compatibilism, which I’ve called ‘the control argument’. The upshot is that incompatibilists do not need to appeal to manipulation arguments to show that compatibilism is false, and they thereby have a second way to avoid the ‘no generalisation’ objection.

So, I have shown that incompatibilists have two ways to avoid the ‘no generalisation’ objection. The first requires holding that compatibilism is false because causally determined agents are covertly controlled by events external to them. The second way, on the other hand, only requires holding that compatibilism is false because deterministic laws entail that determined individuals could be covertly controlled. The second way, then, avoids the conclusion that blind natural forces may covertly control us. I won’t take a stand on which of these two strategies I think incompatibilists should take. After all, I’m not an incompatibilist; so I’ll leave deciding that up to them. What both of these arguments do suggest, however, as I’ve already indicated, is that if covert control is the responsibility-undermining factor in control cases, then it seems that indeterministic individuals may also be covertly controlled; therefore if compatibilism is false, source incompatibilism may also be false. I’ll return to this line of argument in §7.5. The primary task of the next two chapters will be to argue that compatibilist can resist control cases (and thus a fortiori manipulation arguments). As I’ve argued control cases are the true threat to compatibilism, I will focus on showing why the remaining control cases are not in fact counter-examples to all plausible current compatibilist conditions on moral responsibility (in particular, the synchronic ownership conditions of my structural-narrative conditions, which I sketched in Chapter 3). It is to this task that I now turn.
Chapter 6: Hard Compatibilism and Threatening Covert Control

6.1 Introduction

There is something – for want of a better word – *creepy* about the possibility of covert control.¹²⁶

We don’t like the idea that someone or something secretly controls what we do. Source incompatibilists – in a manner similar to historicists – appeal to control cases to support their position; they claim that an individual who satisfies *their* conditions on moral responsibility cannot be covertly controlled, and therefore is morally responsible for her actions. In this chapter and the next, however, I argue that my version of *hard compatibilism* – the thesis that covertly controlled (and causally determined) individuals may be morally responsible for their actions – offers the most stable and plausible response to control cases.

While the previous chapter was concerned primarily with structural issues about manipulation *arguments*, this and the next chapter focus on control *cases* – that is, the apparent counter-examples to all plausible current compatibilist conditions on moral responsibility that support control and manipulation arguments. My aim is to show that none of these cases are counter-examples to (at least) my compatibilist conditions on moral responsibility; hence these cases do not undermine compatibilism.

My general response to control cases, as noted, will be the stance of hard compatibilism. Unlike Frankfurt (the most notorious hard compatibilist), I seek to explain the deeper reason behind the hard compatibilist’s claim that covert control is not necessarily responsibility-undermining. In §6.2, I distinguish unthreatening from threatening control cases, and then I defend different (though related) responses to those different sorts of cases. As I argue in §6.2, the distinction between these two sorts of cases depends on intrinsic features of those cases.

In this chapter, I respond to the remaining threatening cases. Here I extend my strategy from Chapter 4. There I argued brainwashing cases are not counter-examples because a post-

¹²⁶ This is how Tognazzini (2014: 358) describes manipulation; it seems much more fitting with respect to covert control. I make clear the difference between ‘manipulation’ and ‘covert control’ in §6.2.
brainwashing individual is a different narrative self to the pre-brainwashing individual. In §6.4, I argue that non-sporadic continuous manipulation cases – a variant of Pereboom’s Case 1 – are not counter-examples because continuous manipulation creates a new agent and we have no reason to believe this new agent is not morally responsible. I argue that continuous manipulation only seems responsibility-undermining because we misidentify the locus of moral responsibility in cases featuring such manipulation; once the actual locus is identified, the non-responsibility judgement soon dissipates. I then apply this response to sporadic continuous manipulation cases (e.g. Pereboom’s Case 1). But it does not follow that in such cases a new agent is created. However, I argue that a new narrative self is created as a result of such manipulation, and I argue we have no reason to think this new narrative self is not morally responsible.

Before presenting my response to continuous manipulation cases, I argue in §6.3 that soft compatibilist responses to such cases – the strategy of identifying a relevant difference between a covertly controlled individual and a merely causally determined individual – are unsuccessful. In particular, I argue that Kristin Demetriou’s (2010) response is unsuccessful because while she claims that there are no metaphysically coherent ways to understand the mechanics of continuous manipulation, I show that there is at least one way to do so. This clears the way for my hard compatibilist response to such control cases.

6.2 Threatening and unthreatening covert control

In Chapters 2-4, I discussed hypnosis, brainwashing, and design control cases. The hypnosis case – Hypnotised Harry – featured a hypnotist who induces an individual, Harry, with (for example) a desire and corresponding value.\(^\text{127}\) The value does not cohere with the Harry’s other values. Despite this, Harry acts on the induced desire and expresses an induced value when he acts.

Given that Harry is subject to constraining covert control – that is, covert control which produces

\[^{127}\text{Note that this case can be construed so that it is a counter-example to any of the standard structural conditions on synchronic ownership. I’ve opted to construe it as only a counter-example to Watson’s conditions for the sake of simplicity. As I mentioned in §2.2, I am neutral on which of the standard structuralist conditions is preferable.}\]
an ‘abnormal’ psychological life such that Harry is unable to make sense of his actions (see §3.5) – it seems that Harry is not morally responsible for his actions. Hence, hypnosis cases are counter-examples to the standard structuralist conditions on moral responsibility. I argued that such cases suggest that there is a narrative authenticity condition on synchronic ownership – namely that an individual, when she acts, must form a disposition to provide a narrative explanation for that action, which makes sense of that action with respect to the individual’s character, past actions and/or planned future actions. Given that Harry cannot provide such an explanation for his hypnotised action, on my view, Harry is not morally responsible for that action.

The brainwashing case I discussed – Brainwashed Beth – features manipulators who brainwash Beth, an exceptionally sweet person, to have the system of values of a serial killer. Beth then kills her neighbour, George. In Chapter 4, I argued that Brainwashed Beth is not a counter-example to my structural-narrative view. I argued that narrative selves, and not individuals, are the loci of moral responsibility. Because post-brainwashing Beth/Beth\textsubscript{V2} is a different narrative self to pre-brainwashing Beth/Beth\textsubscript{V1}, there is no worry that Beth\textsubscript{V1} (if she returns) will be morally responsible for Beth\textsubscript{V2}’s actions; thus there is no good reason to think that Beth\textsubscript{V2} is not morally responsible. Hence, it is plausible for compatibilists to accept that Beth\textsubscript{V2} is morally responsible.

One reason I noted why some might think that Beth\textsubscript{V2} is not morally responsible is that she is covertly controlled. Beth\textsubscript{V2} is covertly controlled by her manipulators because they create her to be a specific way in conjunction with their knowledge of her future environment. As Beth\textsubscript{V2}’s manipulators – or, rather, creators – plausibly do have such knowledge of her future environment, it seems that she is covertly controlled.\textsuperscript{128} However, I argued that this sort of covert control is not responsibility-undermining because brainwashing is similar to design, and

\textsuperscript{128} As I noted in §4.8, even if this feature is not stipulated, it is implied.
compatibilists must accept that designed individuals are morally responsible. Design and brainwashing are both creation cases. The only difference is that in design cases the individual exists for longer and during that time develops into a particular sort of person, while a brainwashed individual comes into existence fully formed as a particular sort of person.

In brainwashing and design cases, manipulators must have knowledge of or local control over an individual’s environment in order to successfully covertly control the individual, as I discussed in §2.7 and §5.2. Let’s call these ‘additional factors’ as they go beyond merely interfering with an individual. Hypnosis cases, on the other hand, do not seem to require such additional factors for a manipulator to control her victim. And the same is true with sporadic (and, as we’ll see in §6.3, non-sporadic) continuous manipulation, as featured in Pereboom’s Case 1: even without additional factors, the neuroscientists control Plum1. However, in all other types of control case additional factors are specified and required. For instance, in Designed Ernie, Ernie is designed by Diana. But Diana doesn’t just design him; she designs him in conjunction with her knowledge of all facts about the past and the laws of nature. Only this way is Diana able to ensure that Ernie performs the action she wants him to perform, only this way can Diana control Ernie, and only this way will Ernie seem to be lacking responsibility. While just ‘setting up’ an individual means that an individual has been (in some sense) ‘manipulated’, it does not mean that the individual is covertly controlled. Design, conditioning, programming, and the ego button (as I’ll discuss in §7.2) are simply different ways to ‘set up’ an individual. Incompatibilists are therefore required to include an additional factor to turn such forms of mere manipulation into covert control, and this is why additional factors are included in many control

---

129 Although I classify brainwashing cases as ‘unthreatening’ control cases below, they might not fall neatly into this category. It seems that on some readings brainwashing might count as a threatening control case – i.e. Beth is covertly controlled without additional factors being specified. As I note below, though, the point of this classification is so that I can generalise my response to other unthreatening control cases (e.g. design, programming, etc.) in Chapter 7. Since I have already responded to brainwashing cases in Chapter 4, it’s not a problem that brainwashing cases might not fall neatly into one category.

130 As I argued in §5.4, the term ‘manipulation’ is somewhat imprecise. In the literature, ‘manipulation’ is sometimes used in the sense of simply interfering with an individual – i.e. by manipulating her brain, or some such. Other times, it is used as synonymous with covert control – i.e. manipulating an individual.
cases. Because the manipulation in these cases is not strong enough to count as a mode of covert control without some additional factors, I shall say that such cases feature weak manipulation. I shall call control cases that feature both weak manipulation and additional factors ‘unthreatening’ cases. The main point of this classification is simply so that my response to one form of unthreatening case can be generalised to other unthreatening cases. I respond to unthreatening cases in Chapter 7.

Given that a hypnotised or a continuously manipulated individual can be covertly controlled without it being specified that some additional factors are in play, I shall say that such cases feature a strong mode of manipulation – that is, a mode of manipulation which is sufficient to alone count as a mode of covert control. And I shall call such cases ‘threatening’ control cases. In the remainder of this chapter, I respond to the final modes of threatening covert control: sporadic and non-sporadic continuous manipulation.

6.3 Soft compatibilism vs. sporadic continuous manipulation

Soft compatibilists (e.g. Fischer and Ravizza 1998; Fischer 2004; Mele 2005; Baker 2006; Haji 2009; Demetriou 2010) propose further conditions on moral responsibility which they claim a particular covertly controlled individual has not satisfied. In §3.2, I used this strategy to respond to hypnosis cases. I argued that there is a narrative condition on synchronic ownership that a hypnotised individual does not satisfy; hence hypnosis cases are not counter-examples to my compatibilist conditions on moral responsibility. While I advocated a soft compatibilist reply, I did so, as noted in §3.5 and §6.2, because hypnosis is a form of constraining covert control. I do not, however, advocate soft compatibilism when it comes to nonconstraining covert control.

Historicists respond to brainwashing cases by positing further, namely historical, conditions on moral responsibility.131 Such a strategy, as I argued in §2.7, is unpromising.

131 Positing historical conditions in response to hypnosis cases might seem to work, given that I’ve claimed that hypnosis cases are constraining control cases and that further conditions may be posited to avoid such cases. However, I argued in Chapter 3 that there are plausible structural conditions that a hypnotised individual does not
Brainwashing cases feature nonconstraining covert control, and, as McKenna (2008a: 143-144) points out, incompatibilists therefore have a straightforward response to the historicist: change the case so that the individual satisfies the extra conditions on moral responsibility that a soft compatibilist posits. Incompatibilists can therefore change from one nonconstraining control case to another – for example from a brainwashing case to a design case. The problem for historicists, as I argued in §2.7, is that both are cases of nonconstraining covert control and those cases are structurally analogous, so the historicist cannot plausibly have different judgements about these cases (and the fact they might only highlights the instability of historicism).

Compatibilists must accept that a designed individual can be morally responsible, and that is why I argued that compatibilists must say that a brainwashed individual (or, rather, the brainwashed time-slices of that individual) are also morally responsible.

Historicists, most notably Mele (2005), have also responded to Pereboom’s Case 1 – which features sporadic continuous manipulation (a form of nonconstraining covert control). Mele argues that Plum1 does not satisfy some historical conditions on moral responsibility, so Case 1 does not motivate an argument against compatibilism. Mele writes:

[Plum1] played no role at all in shaping his procedure for weighing reasons (say, through trial and error over the years he has been in the business of deliberating). Unlike normal agents, Plum[1] has no control throughout his history as an agent over this important aspect of his deliberative style. (2005: 78)\textsuperscript{132}

\textsuperscript{132} Note that Mele clearly characterises his historical condition as a positive one here, while he officially endorses a negative one (at least when he is wearing his compatibilist hat). But, as I argued in §2.5, while Mele officially endorses a negative constraint on moral responsibility, he implicitly places a positive historical condition on it. His point here emphasises my earlier point as Mele makes it explicit that he places a positive historical condition on moral responsibility.
Thus Mele claims that Plum1 is not morally responsible because he lacks the sort of control over his development that a ‘normal’ (presumably causally determined) individual has. This apparent relevant difference can then be turned into a historical condition on moral responsibility. While I have argued that historicism is untenable, Pereboom has a different way of avoiding Mele’s response: he changes the case. Pereboom (2008b: 18) claims that we must simply imagine that rather than Plum1 being sporadically continuously manipulated that he is continuously manipulated throughout his entire life. So, rather than being directly controlled like a puppet for various periods of his life, Plum1 is controlled like a puppet throughout his entire life – that is, the neuroscientists provide Plum1 with all his neural/psychological states through his entire life. The idea is that Plum1 is covertly controlled and then kills White while satisfying any historical condition on moral responsibility that Mele – or any other historicist – posits, and it certainly seems that Plum1 satisfies these conditions when he is continuously manipulated. Hence Mele’s historicist reply only delays the historicist defeat.

Soft compatibilism, then, seems unable to respond to continuous manipulation cases. Kristen Demetriou (2010), however, has presented a new sort of soft compatibilist reply to Case 1. She argues that the details of Case 1 are ambiguous and that they are open to several interpretations. She claims that each interpretation can be responded to with either a medium compatibilist reply or a soft compatibilist reply. That is, we can either reject the judgement that Plum1 is not morally responsible by arguing that the judgement is unreliable (by employing the ‘reverse generalisation strategy’ I used to respond to design cases in §4.8) or claim that there is a relevant difference between Plum1 and a merely determined individual. I shall focus on the interpretation of Case 1 that Demetriou claims avoids a medium compatibilist reply, but not her soft compatibilist reply. My reason for doing so will soon become clear.

---

133 Mele’s historical condition is a ‘no bypassing’ condition. This requires that an individual’s reflective control over her mental life is not undermined. Plum1 seems to retain such reflective control; hence Mele’s condition is satisfied in Case 1. One might, then, develop Mele’s negative historical condition into a ‘no covert control’ condition. I noted in §5.3 that a ‘no covert control’ condition seems ad hoc; I discuss this condition again in §7.5.
While other soft compatibilist replies identify a condition on moral responsibility that Plum1 has not met, Demetriou identifies an alleged condition on agency that he has not met. Demetriou points out that a merely determined individual is causally integrated in a particular sort of way – that is to say, if we observe the causal structure of a merely determined individual, we will see that his neural (or brain) states realise particular psychological (or mental) states, and these neural states cause further neural states and those psychological states cause further psychological states. Fig. 1 illustrates this causal integration.

Plum4
\[
\begin{align*}
M1 &\rightarrow M2 &\rightarrow M3 \\
B1 &\rightarrow B2 &\rightarrow B3
\end{align*}
\]

FIG 1. (Demetriou 2010: 607)

Compare this now to Fig. 2, which illustrates Plum1’s apparent causal structure.

Plum1
\[
\begin{align*}
M1 &\rightarrow M2 &\rightarrow M3 \\
B1 &\rightarrow B2 &\rightarrow B3 \\
NS &\rightarrow &\rightarrow NS
\end{align*}
\]

FIG 2. (Demetriou 2010: 607)

Plum4 is a counterpart of Plum1 except that Plum4 has not been subject to what Demetriou (2010: 608) calls ‘suppressive’ manipulation, of which sporadic continuous manipulation is a form. The manipulation is suppressive because it apparently undermines the causal efficacy of Plum1’s prior mental states: his subsequent states are caused, not by his previous states, but by

---

134 How we interpret the details, as Demetriou (2010: 607, n.8) notes, depends on our view about mental causation. What I say in the remainder makes no claims about mental causation, and I remain neutral on this issue throughout.

135 As Demetriou (2010: 607, n.8) notes, these diagrams are usually used to present ‘different visions of mental causation’. Also, Demetriou talks about brain and mental states rather than neural and psychological states as I have been. This difference in terminology is not significant in what follows.
the neuroscientists’ activities. Plum1’s brain and mental states are causally isolated; one brain state realises a single mental state, but his brain and mental states do not cause his subsequent brain and mental states. On the other hand, there are no impediments to Plum4’s brain states causing his subsequent brain states, which realise his subsequent mental states. Although Plum1 and Plum4 have exactly the same mental states and therefore have qualitatively identical mental lives, Demetriou claims that they do not share ‘the same status in terms of agency’ (2010: 608) – that is, while Plum4 is an agent when he kills White, Plum1 is not when ‘he’ kills her.

Cast in this way, Demetriou’s reply is very convincing. She appears to have identified a relevant difference that shows that Case 1 is not a counter-example to all plausible current compatibilist conditions on moral responsibility, and because she claims to have identified a condition on agency that Plum1 has not satisfied, her response can be used by all compatibilists. But Demetriou’s argument supports a rather strong claim. She claims to have shown every metaphysically coherent interpretation of Case 1 succumbs to either a medium compatibilist or soft compatibilist response. To undermine her argument, then, I need only show that there is one metaphysically coherent interpretation of Case 1 that avoids her soft compatibilist response and that avoids a medium compatibilist response. I’ll start by arguing that there is an interpretation of Case 1 that avoids her soft compatibilists response, and then I’ll argue that a medium compatibilist response is prima facie ineffective in response to this interpretation.

The viability of Demetriou’s soft compatibilist response depends on the plausibility of two claims: (1) that the suppression of an individual’s prior states must undermine an individual’s agency, and (2) that there is no causal integration after the neuroscientists’ first intervention (B1/M1 in Fig 2.). I will first argue that (2) is false. While it might be plausible that the neuroscientists’ first intervention (i.e. the induction of B1/M1) suppresses Plum1’s prior states (i.e. those which precede those shown in the diagram), which I’ll grant for the sake of argument,

---

136 She actually calls them ‘hard-line’ and ‘soft-line’ replies, following McKenna (2008a). I have opted for different terminology to make the various compatibilist strategies easier to appreciate – in particular, we might reject the non-responsibility judgement (i.e. M1 of MAS) in more ways than McKenna considers.
there is a plausible understanding of the manipulation such that there is causal integration after the first intervention (i.e. there is causal process running between B1/M1-B2/M2-B3/M3). I will then argue that (1) is false. There are lots of instances in real life where an event might ‘suppress’ prior states without an individual’s agency being undermined (such as those involving motivational posters).

The diagrams that are pivotal to Demetriou’s argument are normally used to display the relationship between brain states and mental states (see n.135). (In what follows I will talk in terms of brain and mental states to cohere with the diagrams that I discuss.) They do not accurately portray the causal structure of the human brain/mind. The causal workings of the human brain are clearly much more complicated that the diagrams suggest. In particular, Demetriou’s diagrams suggest that normally a particular brain/mental state is sufficient for the production of a subsequent brain/mental state. But it is not the case that a particular brain/mental states is alone sufficient for the production of a subsequent brain/mental state; there are lots of background conditions that must also be satisfied. This is in much the same way that the impact of a billiard ball A is not alone sufficient to move a billiard ball B; it also matters what force A was moved with, what the surface the balls are on is like, what forces (e.g. gravity) are in play, what mass the billiard balls are, and a whole host of other background conditions. The same is true with causation between states. For a particular brain/mental state to produce another brain/mental state, certain background conditions must be satisfied. One of the most important of these background conditions is often the external world.

It is uncontroversial that our reasoning processes are regularly influenced by the external world. Perceiving things in the world might cause us to consider certain options or it might help us to make a decision we had been struggling to make. For example, a motivational poster that asserts that we should just do it might make us think about the things we can do (when previously we might not have considered these options) or it might help us to decide to do something we had been deliberating about. Such conscious influence on our reasoning processes is common.
There is also substantial empirical evidence that the world can unconsciously affect our reasoning processes. For example, prisons have often been painted Baker-Miller Pink in an attempt to reduce aggression among inmates; several studies have supported the hypothesis that this is effective.\textsuperscript{137} Studies have also suggested that smells and sounds can affect our behaviour without us noticing.\textsuperscript{138} Hence the conscious and unconscious causal role of the world in the production of actions is something we (and any adequate account of agency) must acknowledge.\textsuperscript{139}

As the world has a causal role in the production of action, external events are often important causal contributors (that is, jointly sufficient with our states and other background conditions) in the production of our subsequent states and behaviour. Suppose I feel hungry (B1/M1), so I look in my fridge and I see a cake (B2/M2). I then form the belief that there is a cake in my fridge (B3/M3). As I am hungry and I desire not to be hungry, I make the decision to eat the cake (B4/M4). I then eat the cake. Clearly, there being cake in my fridge is integral to me forming the decision to eat the cake. If there were no cake in the fridge (and I wasn’t hallucinating), then I would not have decided to eat the cake. Demetriou’s diagrams, however, make no room for the causal influence of the external world. Once we acknowledge this, it becomes plausible that the neuroscientists do not have to undermine the causal efficacy of Plum1’s states after their first intervention in order to covertly control him from moment to moment.

Let’s think about Case 1 again. According to Demetriou’s interpretation of the mechanics of the manipulation, each of the neuroscientists’ interventions suppresses a prior state of Plum1’s. So when B1/M1 is induced this suppresses whatever Plum1’s prior brain/mental state was by rendering those states causally impotent, when B2/M2 is induced this suppresses B1/M1, when B3/M3 is induced it suppresses B2/M2, and so on. In other words, there is no

\textsuperscript{138} Baron (1997).
\textsuperscript{139} The UK government also has a department – affectionately known as the ‘nudge squad’ – which works on improving people’s behaviour in subtle ways. In other words, this department seeks to manipulate people. See https://www.gov.uk/government/organisations/behavioural-insights-team/about for more details.
causal process running between B1/M1-B3/M3. But that’s a really strange thing for the neuroscientists to do. They, after all, are trying to covertly control Plum1 so that he kills White; they are trying to do so by inducing his brain/mental states so that he engages in an egoistic reasoning process which will lead him to killing White. Now if this is their aim, why would each intervention (after the first intervention) necessarily undermine the causal process between B1/M1-B3/M3? That seems somewhat counter-productive. Presumably the initial state they induce could easily play a role in producing the subsequent states the neuroscientists wish Plum1 to have. The neuroscientists would then only have to intervene to the extent that B1/M1 on its own is not sufficient for the production of B2/M2. Thus, after the first intervention the neuroscientists need only act as a kind of *causal enforcer* – that is, the neuroscientists’ activities would not be sufficient for B2/M2 to obtain, but their activities would provide whatever extra is required for B2/M2 to obtain. The neuroscientists are, in a sense, causing B2/M2, but they are not causing it alone; B1/M1 just has a more important role to play.

This is similar to how a friend might continue to badger us to do something even after we have agreed to do that something. While there are lots of times our friend’s badgering might overdetermine (i.e. we will do the thing regardless), pre-empt (i.e. we would have done the thing, but the badgering is sufficient to make us do it independently), or even be causally irrelevant (i.e. the badgering has no effect on us) to the thing we’ve agreed to do. There are other times when we want to do something, but we will only actually do it if we are badgered. In such cases, the badgering alone is not sufficient for us to act; rather, it is only sufficient in conjunction with our prior states.

After the first intervention, the neuroscientists need only match the causal influence of such a friend in the latter sort of case. The neuroscientists would still be controlling Plum1 ‘moment to moment’ because they are still controlling exactly what his states are like from moment to moment; it’s just that they have less work to do after the first intervention, which does much of the heavy lifting. In the same way, if a friend thinks she has convinced us to do
something (and she is smart enough to avoid being causally irrelevant, overdetermining or preempting us) she has less work to do to keep us on track. But she might still have some work to do because we might change our minds.

Of course, our friend would only realise she had more work to do if it seemed like we might change our minds. She might be able to judge that this might happen by observing our change in body language. Our body language (among other things), then, acts as a kind of feedback mechanism. So, for the neuroscientists to covertly control Plum1, there must be some sort of feedback mechanism between Plum1’s brain and the neuroscientists. That is, the neuroscientists must have some way of knowing whether the state they have just induced will be sufficient (in conjunction with background factors) to produce the subsequent state they wish Plum1 to have. When a prior state is not sufficient, the neuroscientists can, so to speak, make up the difference and help to produce the subsequent state. But, note, the initial state they induce will have a role to play in the production of Plum1’s subsequent state. FIG. 3 displays this feedback mechanism.

![Feedback Mechanism Diagram]

The arrow between the neuroscientists is there to indicate that there is some causal process obtaining between the neuroscientists’ activities. The arrow upwards from the neuroscientists to the brain states indicates their interventions into Plum1’s brain. The arrow downwards from the brain states to the neuroscientist is the feedback mechanism. This allows them to tailor their subsequent intervention as required. Once the feedback mechanism is in place, of course, there no longer seems to be any problem with claiming that B1/M1 causes B2/M2 and that B2/M2

FIG. 3 displays this feedback mechanism.
causes B3/M3. It is not as if normally brain/mental states alone are sufficient to produce subsequent brain/mental states; there are always background conditions which are jointly sufficient with the prior brain/mental state. These are overlooked in discussions about mental causation because our interest is in the relationship between brain and mental states, and not on how brain states are produced.

As I’ve interpreted Case 1, it sounds similar to the case Demetriou (2010: 604-605) considers, where the neuroscientists and Plum1 are joint sufficient causes of Plum1’s subsequent states. Even in such a case, she claims that the neuroscientists would have to ‘undermine the causal efficacy of Plum1’s states in some way’ because, presumably, Plum1 is the sort of individual who would act without the neuroscientists’ intervention, and hence his states (without intervention) are normally causally sufficient to produce his subsequent states. So, according to Demetriou, the neuroscientists would have to undermine Plum1’s states to some extent, but not entirely, so that those states in conjunction with the neuroscientists’ input would produce Plum1’s subsequent actions. Moreover, she claims that their interventions would be ‘faithful’ or ‘unfaithful’. If they intervene faithfully, then they cause Plum1 to act as he would have acted without their intervention. If they intervene unfaithfully, then they cause Plum1 to act differently. Demetriou claims the faithful interpretation is open to a medium compatibilist response, and the unfaithful response is open a soft compatibilist response. While the case I am sketching is more like an unfaithful intervention (which seems necessary to elicit the non-responsibility judgement), there are important differences between the cases.

First, I have accepted for sake of argument that the neuroscientists’ first intervention entirely undermines the causal efficacy of Plum1’s prior states (that is, his states prior to B1/M1), so B1/M1 is produced solely by the neuroscientists’ efforts, whereas in Demetriou’s joint sufficiency case Plum1 and the neuroscientists are joint sufficient causes of B1/M1 – that is, in Demetriou’s case, Plum1’s prior states are not rendered entirely causally impotent; the neuroscientists simply undermine those states by (somehow) reducing their causal efficacy by
(somehow) making those states not quite sufficient for the production of subsequent states.

Second, I deny that the neuroscientists have to undermine the causal efficacy of Plum1’s states after the first intervention – that is, I deny that the neuroscientists reduce the causal efficacy of B1/M1, B2/M2, and B3/M3. After all, it seems plausible that the neuroscientists induced B1/M1 because it would have a causal role in the production of Plum1’s subsequent states. So the neuroscientists do not need to reduce the causal efficacy of B1/M1 for B2/M2 to come about, nor do they need to reduce the causal efficacy of B2/M2 for B3/M3 to come about. The neuroscientists can, instead, act as causal enforcers, adding whatever is required (if anything is required) for B1/M1 to produce B2/M2, and for B2/M2 to produce B3/M3 in the same way that the world would have to be a particular way for any individual’s brain/mental states to cause a subsequent brain/mental state. According to Demetriou, if Plum1’s states are not causally efficacious then he does not exert causal control over his actions. But even if B1/M1 is not alone sufficient for B2/M2 to obtain, that doesn’t mean that B1/M1 is not causally efficacious.

The important thing to realise, and what I have been trying to draw attention to, is that there is much more to the causal story than simply one brain state causing another brain state. If causal control required that brain/mental states alone had to be sufficient for subsequent brain/mental states to obtain, then no one would ever exert causal control because there are always background necessary conditions. We usually ignore these background conditions because they are not the focus of our investigation (e.g. when we are investigating mental causation we are only interested in the relationship between brain and mental states). As I argued earlier, the external world (in many different ways) is a background condition, and the external world and our states are jointly sufficient for the production of our subsequent states. But once we include the external world in the causal story, we do not conclude that we are not agents. The world, we might say, is a causal enforcer: it ensures that our states are causally sufficient for the production of our subsequent states. The neuroscientists, then, simply take on some of the role of the world: they ensure that Plum1’s states are causally sufficient for the production of his subsequent states.
Just as we have no reason to think that the world suppresses our agency, we have no reason to think that the neuroscientists, on the version of Case 1 I have sketched, undermine Plum1’s agency after their initial intervention. Hence, even if the neuroscientists’ initial intervention suppresses Plum1’s prior states, this does not rule out there being a causal process running between Plum1’s subsequent states (i.e. between B1/M1-B3/M3).

If there is a causal process between his states, then when Plum1 kills White he is an agent. If she wished to maintain that Plum1 is not an agent, Demetriou would have to argue that there is something significant about the initial ‘suppression’ of Plum1’s states prior to the manipulation period. But this claim is dubious. There are lots of instances in real life where our states are ‘suppressed’, according to Demetriou’s definition, to make way for new ones. For instance, suppose I’m thinking about the Four-Case Argument and suddenly I hear my window being smashed. The event of the window being smashed stops me from thinking about the Four-Case Argument and makes me think about my window being smashed. Rather than thinking, ‘how is the inference to the best explanation supposed to work?’ I am suddenly thinking, ‘why is my window being smashed? Is someone trying to break into my house?’ Etc. It looks as if the event of my window being smashed suppresses my prior states in the sense that Demetriou claims undermines agency. According to her, something suppresses a state if it undermines the causal efficacy of that state. This seems to happen when my window is smashed. If my window had not been smashed, I would have continued thinking about the Four-Case Argument. But as my window has been smashed, my thoughts change. Hence, the causal efficacy of the brain/mental states that underpinned my thinking about the Four-Case Argument are suppressed by the event of my window being smashed. But it is implausible to suppose that hearing my window being smashed undermines my agency because events like this occur regularly in real life. It seems that any plausible account of agency must accommodate such a regular occurrence.

Perhaps there is some significant difference between Plum1 and me which explains why I am an agent and Plum1 is not. But the only difference between us is the method of the causal
suppression of our prior states, and that seems like an ad hoc difference in this context. In any case, if there were some non-ad hoc difference, incompatibilists could just change the case by modifying the mode of manipulation the neuroscientists use to covertly control Plum1; if the problem is that the manipulation is ‘inside’ the brain, the neuroscientists need only move the manipulation to ‘outside’ the brain. They might hire a biochemist to devise a poison or series of poisons and a sound technician to devise a series of subsonic sounds to covertly control Plum1. The poison might initiate an egoistic reasoning process and then the subsonic sounds could be applied as a causal enforcer if it ever looks like Plum1 needs some help to keep reasoning egoistically. The neuroscientists could therefore achieve the same effects as radio wave brain manipulation without any worries about Plum1’s agency being suppressed.

But incompatibilists don’t need to change this case because there isn’t a non-ad hoc difference between a state being suppressed by a radio wave and a state being suppressed by a sound (e.g. the sound of a window being smashed). Demetriou’s reply only seems convincing because we overlooked the causal role of the world in the production of action and the possibility of a feedback mechanism between Plum1 and the neuroscientists. Once we see the causal role of the world and include a feedback mechanism, we can understand the causal role of the neuroscientists after their first intervention such that they do not undermine Plum1’s agency. Hence, it seems that there is an interpretation of Case 1 which avoids Demetriou’s soft compatibilist response.

What about a medium compatibilist reply – that is, one that uses the reverse generalisation strategy? I used this strategy to respond to brainwashing and design cases in §4.8. According to this strategy, we must start with our judgements about merely determined individuals and then generalise them to covertly controlled individuals. For this reply to be successful, it must be the case that we can generalise our judgement about a merely determined individual to Plum1, who is a victim of sporadic continuous manipulation. Even as a committed (hard) compatibilists, I find that my judgement that (say) Plum4 is morally responsible does not
generalise to Plum1. There is something particularly worrying – or creepy – about the mode of covert control that Plum1 is a victim of such that I cannot accept he is morally responsible just because he is apparently no different from Plum4. Therefore, Demetriou’s overall argument against Case 1 is unsuccessful because there is an interpretation of Case 1 which avoids both her soft compatibilist response and a medium compatibilist response.

6.4 Hard compatibilism to the rescue

Case 1, then, still poses a problem for compatibilists. In much the same way that I argued that brainwashing leads to the creation of a new moral agent, I will argue that non-sporadic continuous manipulation leads to the creation of a new agent, and that this new agent may then be morally responsible for her actions. I will then argue that there are two interpretations of Case 1, a sporadic continuous manipulation case. On the first interpretation, a new narrative self is created analogous to how a new agent is created in non-sporadic continuous manipulation cases. On the second, no new narrative self is created, but I argue Plum1 is morally responsible. To set things up, I will modify Case 1, in the way that Pereboom does to evade Mele’s (2005) critique of the Four-Case Argument, to become a non-sporadic continuous manipulation case. As I mentioned in §6.3, Pereboom writes:

One might specify that the manipulation takes places at every moment, and directly affects Plum[1] at the neural level, with the result that his mental states and actions feature the psychological regularities and counterfactual dependencies that those of an ordinary agent might exhibit. (Pereboom 2008b: 18; my emphasis)

While it is a bit unclear what the temporal scope of ‘at every moment’ is, let’s suppose that it means at every moment of Plum1’s life. So, rather than being sporadically continuously manipulated, Plum1 is continuously manipulated over every moment of his existence. Let’s call
this Plum1 ‘Plum1c’ and this case that features this ‘Case 1c’. In this case, the neuroscientists use advanced radio technology to induce every one of Plum1c’s brain states, which in turn realise all his mental states throughout his life. So Plum1c’s entire mental life is the result of neuroscientists’ interventions. Given this, it seems that Plum1c is not morally responsible because he is manipulated in the most worrying and creepy way imaginable – viz. he is a literal puppet of the neuroscientists as they control every single thought he ever has or action he ever performs. But, just as Plum1 seems to satisfy the compatibilist conditions on moral responsibility, it seems that Plum1c satisfies them too.

I think that it seems so plausible that Plum1c is not morally responsible because we implicitly identify him with his human animal (call it Bio-Plum). Bio-Plum is clearly not an agent in this case because it doesn’t make any decisions. So, Bio-Plum is not morally responsible for killing White. This seems correct. But there is an agent in this case; an agent who decides to kill White, and who satisfies the compatibilist conditions on moral responsibility in doing so. But who is this agent? My contention is that the agent in this case is Neuro-Plum – an agent composed by Bio-Plum and the neuroscientists; it is Neuro-Plum who is morally responsible for killing White.

This contention might seem implausible. How can a group of agents and a human animal compose another agent? To make this contention more plausible, I will start by taking full advantage of Pereboom’s (2001: 115-116) claim that we can replace the neuroscientists with an intentionless machine. It seems much more plausible that an agent may be composed of a machine and a human animal. For example, suppose that Bert discovers that he has a degenerative neurological condition. A doctor offers to remove the degenerative part of his brain and to replace it with a machine that performs all the functions of that part of Bert’s brain. It seems that Bert would still be an agent, just one that is partially composed by a machine. Moreover, we could imagine that Bert slowly had his brain (bar his brainstem so that he continues to be the same human animal) replaced with a machine that matched all his brain’s functions. It seems that Bert would still be an agent after all these procedures; he would just be
one that was composed by a human animal and a machine. This machine need not even be in Bert’s brain; it could be locked in a safe somewhere and linked into Bert’s nervous system wirelessly.  

Let’s suppose that Plum1_{C} (that is, Bio-Plum) is continuously manipulated by a machine that matches the functions of the neuroscientists, and let’s call this machine ‘Neurobot’. Neurobot is not an agent; it is not conscious, it has no intentions, and it has no intentional states. To avoid the worry that Neurobot has been designed by an agent, we should also, as Pereboom (2001: 115-116) proposes, imagine that Neurobot comes into existence spontaneously without design. Moments after coming into existence, Neurobot induces Plum1_{C} with all his initial neural states, and then with all his neural states throughout his entire life; so all of Plum1_{C}’s psychological states are realised by neural states induced by Neurobot. The agent composed of Neurobot and Plum1_{C} is Neuro-Plum.

Is Neuro-Plum morally responsible for killing White? There is no good reason to think not. Plum1_{C} seemed to lack moral responsibility because he was covertly controlled by the neuroscientists. However, Neuro-Plum is composed by the thing which allegedly controls him; in effect, Neurobot acts as a kind of prosthetic brain. So it seems false to say that Neuro-Plum is covertly controlled by Neurobot, in much the same way that it seems false to say that a brain covertly controls a human person. It seems false because it suggests that normally there is some further thing (such as a soul or Cartesian Ego) that the person is, and if the brain controls the person then this further thing doesn’t get a look in. But we haven’t got any reason to believe there is such a further thing, so we have no reason to believe the brain controls the person. While controversial, it is more accurate to think of a brain (or the psychological states it houses) as being, constituting, or realising the mental states that constitute the person. So the activities of the brain are part of the person’s control over herself, and not something that threatens a person’s control.

---

140 Dennett (1978) proposes a case similar to this. In Dennett’s case, an individual has his brain removed from his head and it is connected to the rest of his body via radio. Valdman (2010: 777-779) also has a similar case.
Given that Neuro-Plum is not controlled by Neurobot, there is no reason not to think he is morally responsible for killing White; hence we should think he is morally responsible for killing her.

This undermines Pereboom’s claim that we can replace the neuroscientists with machines whilst maintaining the non-responsibility judgement. It seems clear that once we replace the neuroscientists with machines in Case 1c that the agent in this case – Neuro-Plum – is in fact morally responsible (or, at least, we have no reason to think he is not). Remember that Pereboom’s argument for the ‘no difference’ premise of the Four-Case Argument (i.e. 4CA-MAS) requires that it is plausible that agent-manipulators are not the main factor in the production of the non-responsibility judgement, and to show this Pereboom claims the non-responsibility judgement is not affected by replacing agent-manipulators with a machine. But I have shown that once we replace the neuroscientists with a machine and we spell out the case, the non-responsibility judgement vanishes because the ‘manipulator’ is actually part of the agent. Hence, there is no argument for the ‘no difference’ premise of the Four-Case Argument, and so the Four-Case Argument (supported by Case 1c) is unsuccessful.

But this still leaves a counter-example to all plausible current compatibilist conditions on moral responsibility – Case 1c itself (as opposed to the case where the neuroscientists are replaced with a machine) which I’ve argued is insufficient to show that compatibilism is false. But Case 1c is not a counter-example. If it is acceptable that Neuro-Plum is morally responsible when he is composed of Neurobot and Bio-Plum, then why shouldn’t Neuro-Plum be morally responsible if he is composed of the neuroscientists and Bio-Plum? The only difference is that Neuro-Plum is composed of a group of agents and a human animal, instead of a machine and a human animal.

Think about Bert again. Suppose that the doctor, who slowly replaced Bert’s brain with a machine, lied to Bert about how the machine works. Suppose that the machine is, in fact, a radio transmitter hooked into his nervous system, and rather than the machine matching the function
of Bert’s brain, it is a team of agents in a control centre somewhere. By hypothesis, the team of
gents do all the same work as Bert’s brain-machine; thus it seems that Bert would still be an
agent, except that he is partially composed by a team of agents, rather than a machine. Likewise,
whether Neuro-Plum’s neural states are induced by a machine or a set of agents doesn’t matter.
So, if all Neuro-Plum’s neural states are induced by agents, it follows that those agents are part of
Neuro-Plum.\textsuperscript{141} Hence, there is no reason not to think that Neuro-Plum \textit{is} morally responsible.
Therefore, Case 1\textsubscript{c} is not a counter-example to all plausible current compatibilist conditions on
moral responsibility.

But what about Case 1 – that is, the case where Plum\textsubscript{1} is \textit{sporadically continuously}
manipulated? There are two possible interpretations of this case. On the first, Plum\textsubscript{1} only
reasons egoistically because of the neuroscientists – that is to say, when he is not being covertly
controlled he reasons non-egoistically. On the second, Plum\textsubscript{1} reasons egoistically whether or not
he is being covertly controlled by the neuroscientists. Let’s start with the first interpretation.

On this version of Case 1, Plum\textsubscript{1} normally reasons non-egoistically but whenever the
neuroscientists manipulate him, he reasons egoistically. If it is plausible that continuous
manipulation creates a new agent, then perhaps sporadic continuous manipulation would do so
too but only during the active manipulation period. Of course, I argued in §6.3 that in Case 1 the
neuroscientists are able to covertly control Plum\textsubscript{1} \textit{without} undermining his agency. I argued that
the neuroscientists covertly control Plum\textsubscript{1} such that they do not undermine the causal efficacy of
his states after their first intervention. And I argued that their first intervention, even if it
undermines the causal efficacy of his prior states, does not undermine Plum\textsubscript{1}’s agency, as it only
involves the kind of causal suppression which is a regular occurrence in real life and which we
don’t think suppresses our agency. Thus it seems that no new agent arises in this version of Case
1. However, even if we accept that no new agent arises in cases of sporadic continuous

\textsuperscript{141} Neuro-Plum might be similar to a group agent. Holding Neuro-Plum morally responsible might be similar to
how group agents are held morally responsible. See Rovane (2005) for more on group agents.
manipulation, this does not mean that a new narrative self does not arise. Remember that in this version of Case 1 we are to suppose that Plum1 only reasons egoistically when he is covertly controlled by the neuroscientists. So the neuroscientists’ first intervention (on any occasion) always suppresses the states of an individual who reasons non-egoistically with those of an egoistic individual. We must discover, then, whether Plum1 and Ego-Plum – the time-slices of Plum1 who reason egoistically – are the same narrative self.

There is a good case to be made that Plum1 and Ego-Plum are not the same narrative self – that is to say, they are not narrative connected. Plum1 and Ego-Plum presumably have (for example) different values. While Plum1 has the values of a non-egoistic individual, Ego-Plum has the values of an egoistic individual. Given this, it seems that Plum1 would not be able to provide a genuine narrative explanation of Ego-Plum’s actions – that is, Plum1 would not be able to tell an intelligible story about how killing White makes sense with respect to his character, his past actions, and/or his plans for the future. At most, Plum1 could remember what Ego-Plum would think about all this. But remembering someone else’s narrative explanation and providing your own are two different things. Doing the former does not require a specific dispositional property, whereas doing the latter does. Given that there is a suppression of Plum1’s states (which presumably partly constitute his non-egoistic values and character) with Ego-Plum’s states (which presumably partly constitutes his egoistic values and character), it seems that Plum1 does not have the values (and perhaps other psychological states) that underpin Ego-Plum’s disposition to provide a narrative explanation of his (Ego-Plum’s) actions; hence any story that Plum1 provides for Ego-Plum’s actions will not be a genuine narrative explanation. So, Plum1 does not satisfy the diachronic ownership condition of my structural-narrative view, and therefore is not the same narrative self as Ego-Plum. Hence Plum1 is not morally responsible for Ego-Plum’s actions, and so we have no reason not to think that Ego-Plum is morally responsible for killing White. Again, it seems that the non-responsibility judgement depends on our misidentification of the locus of moral responsibility in this case.
What about the second interpretation? On this interpretation, I contend that the manipulation does not lead to creation of a new narrative self – that is, there is no Ego-Plum in this interpretation. This version of Case 1 is, in fact, quite like the Ego Button case, which I discuss in §7.2. In the Ego Button, an individual is covertly controlled to act in a way that they are pre-disposed to act, and the same thing happens in this version of Case 1. The only difference is that the neuroscientists play a more active role in producing Plum1’s reasoning process than they do in the Ego Button (as we’ll see). But the neuroscientists only play a role equivalent to events in the world in the production of Plum1’s actions – that is, they are aiding in the production of an egoistic agent’s actions. Given this, it seems that a medium compatibilist reply would be effective in response. That is, it seems that my judgement that Plum4 is morally responsible would generalise to this Plum1, thereby rendering the non-responsibility judgement about this Plum1 unreliable. The fact that the neuroscientists play a role in the production of his action is incidental – other events might have easily played this role. Plum1 might be covertly controlled by the neuroscientists, but any non-responsibility judgement that arises from this fact has been shown to be unreliable.

But what about ‘mid-way’ cases where, for example, Plum1 (when he is not being manipulated) sometimes reasons egoistically and sometimes reasons non-egoistically? Such cases might seem problematic for my analysis. But I contend that such cases should be interpreted like the second interpretation of Case 1 – that is, where no new narrative self is created. If Plum1 is disposed to any egoistic reasoning whatsoever, then it seems that the reverse generalisation strategy will be effective in response to such cases. I contend sporadic continuous manipulation seems immune to the reverse generalisation strategy, as I claimed it is in §6.3, only because it seems to imply that an individual is being covertly controlled to act differently from how they would act if they were not being controlled, and this feature is missing from ‘mid-way’ versions of Case 1 and the version of Case 1 where Plum1 reasons egoistically whether or not he is being covertly controlled. Thus compatibilists can accept that any egoistic Plum1 is morally responsible
for killing White because the reverse generalisation strategy undercuts the non-responsibility judgement about such Plum1s.

Case 1, then, however it is interpreted, is not a counter-example to all plausible current compatibilist conditions on moral responsibility. Continuous manipulation – whether it is sporadic or not – cannot be used as part of an argument against compatibilism.
Chapter 7: Hard Compatibilism and Unthreatening Covert Control

7.1 Introduction

In this final chapter, I respond to unthreatening covert control cases. In §7.2 I explicitly endorse the medium compatibilist strategy (e.g. McKenna 2008a; Fischer 2011), which I’ve already implicitly endorsed in response to other control cases (e.g. in response to design cases in §4.8), of employing the reverse generalisation strategy to show that the non-responsibility judgement about a particular covertly controlled individual is unreliable in response to unthreatening control cases. However, in §7.3, I argue that medium compatibilism does not go far enough. I claim that compatibilists should accept that part of the incompatibilist’s strong conclusion to manipulation arguments is correct – namely that all causally determined individuals are covertly controlled by external events. This helps to explain why (natural) compatibilists have the judgements they do, and it also sets up my argument against source incompatibilism, which I sketch in §7.5.

In §7.4, I then argue that hard compatibilism is not as costly as it might seem. I do this by endorsing and defending Paul Russell’s (2010) selective hard compatibilism in response to Patrick Todd’s (2012) argument that compatibilists must erroneously accept that manipulators can blame their victims. Russell argues that we judge that manipulated individuals are not morally responsible because we put ourselves in the position of the manipulators when assessing control cases. Since we rightly deem it unfair to hold an individual morally responsible for an action we have covertly controlled her to perform, we infer that it is unfair to hold them responsible for covertly controlled actions generally. But Russell claims that while others may justifiably hold a covertly controlled individual to account for her actions, the individual’s covert controller cannot. Compatibilists can therefore avoid the counter-intuitive implication which Todd claims shows that compatibilism is false, and this thereby undercuts Todd’s argument. I then defend Russell’s argument from two objections of Pereboom’s (2014), which allows me to explain some of the more nuanced elements of (hard) compatibilism; this helps to increase (hard) compatibilism’s plausibility, and thus decrease its apparent theoretical costliness.
Finally in §7.5, I argue that hard compatibilism is preferable to source incompatibilism. I argue that source incompatibilism collapses into impossibilism in much the same way as I argued that historicism collapses into source incompatibilism in §2.7. Control cases are used to motivate source incompatibilism. But, as I argue, if there are control counter-examples to compatibilism then there are also control counter-examples to the source incompatibilist (i.e. libertarian) conditions on moral responsibility. However, source incompatibilists cannot bite the bullet when faced with these cases; I argue that this results in source incompatibilism collapsing into impossibilism. Hence, if we believe that free will and moral responsibility are possible then we must be hard compatibilists.

In §7.6, I conclude this chapter and this thesis.

7.2 Hard compatibilism vs. unthreatening covert control

An implication of causal determinism is that our actions ultimately have causal sources in events external to us. Control cases attempt to show us that if our actions have causal sources external to us, then we are not morally responsible for those actions. Threatening control cases do this in an especially creepy way. But, following my arguments against threatening control cases, it seems that the creepiness stems from ambiguities or buried details in those cases, and once we clarify or identify those details (e.g. misidentifying the proper agent/narrative self in that case), such cases no longer seem to be counter-examples. Unthreatening covert control cases, however, seem much less creepy; in fact they just seem to describe what is true for any causally determined individual – namely that our actions have causal sources external to us. They do, however, add something creepy: namely that this causal source might be another agent or include other agents.

For compatibilism to have any plausibility, I think that compatibilists must get comfortable with the fact that there will be some instances in which an individual satisfies the compatibilist conditions on moral responsibility despite being covertly controlled by something else. In this section, I argue that compatibilists can accept that unthreateningly covertly
controlled individuals are morally responsible because the non-responsibility judgement about those individuals is unreliable. I will start by discussing a control case of Shabo’s (2010), and then I will extend the response to that control case to other unthreatening control cases (e.g. design, programming, and conditioning cases).

Shabo’s case uses a weak variety of local manipulation. Here is his case:

*The Ego Button.* Natasha is an ordinary deterministic agent, who exhibits something like the usual mesh of egoistic and altruistic tendencies, when she is presented with a remarkable opportunity. With a judicious leak to a local journalist, she can eliminate a nettlesome political rival (one who, for what it is worth, embodies a similar mesh of egoistic and altruistic tendencies), while displacing blame onto a third [party], whom she has always mistrusted. What Natasha doesn’t realize is that a team of powerful neuroscientists has worked behind the scenes to arrange this opportunity, which it will not see slip away. While the neuroscientists would prefer that she deliberate and act on her own, the preliminary neural indications do not look good; and so, as the fated hour approaches, they press the Ego Button. It works. By ramping up activity in one region of Natasha’s brain while suppressing it in another, the Ego Button ensures that her reasoning about the situation will be structured around the question, “Which of my options will best further my interests?” The result is that the situation’s egoistic implications become salient, while its moral implications recede to the margins of her awareness. In this mindset, there is no question that going ahead with the leak is the right move. (Shabo 2010: 176)

Rather than manipulating an individual to act in a way counter to her character, the manipulators covertly control Natasha to act in accordance with her character. As Shabo says, Natasha ‘exhibits something like the usual mesh of egoistic and altruistic tendencies’, so it is fair to say that
Natasha has some egoistic values (e.g. she values getting what she wants, not worrying about the feelings of others, etc.) or higher-order desires.\textsuperscript{142} Is Natasha morally responsible for leaking the story? Shabo claims that she is not because she is not the ultimate source of her action. Let’s assume that Shabo’s claim that Natasha is not morally responsible is correct. Before continuing, I want to dispute his explanation of why she is not morally responsible.

If not being the ultimate source of one’s actions were a sufficient condition on not being morally responsible, we should be able to redescribe Shabo’s case \textit{without} the additional factors (such as the neuroscientist’s knowledge and local control over her environment) while preserving the non-responsibility judgement. But, as I will now argue, we cannot. Let’s suppose that rather than the neuroscientists pressing the ego button, Natasha’s egoistic reasoning processes are ‘ramped up’ by a feature of the world, such as a motivational poster. Suppose that as Natasha is reasoning about whether to leak the story she notices a poster that says \textit{just do it}. This poster matches the causal role of the ego button being pressed. Natasha then leaks the story. Is Natasha morally responsible for leaking the story? It seems so, because there doesn’t seem to be anything responsibility-undermining about an individual’s reasoning processes being causally influenced by a feature of her surroundings. But, by hypothesis, Natasha is not the ultimate source of her action in this alternative scenario; so it seems that not being the ultimate source of an action is not sufficient for non-responsibility.\textsuperscript{143}

How might we explain the difference in judgements between the covertly controlled Natasha and the merely manipulated (by the motivational poster) Natasha? Well, the only difference between the two Natashas is that the former Natasha is covertly controlled, whereas the latter is not. Notice that the neuroscientists satisfy my two criteria for unthreatening covert control: (i) they set Natasha up by ‘ramping up’ her egoistic reasoning processes, and (ii) they

\textsuperscript{142} If Natasha were covertly controlled to do something counter to her character, then this case would be just like a hypnosis case; my response to hypnosis cases would therefore apply here too.

\textsuperscript{143} Tognazzini (2014: 367) also argues that we judge that manipulated/covertly controlled agents are not morally responsible because they are not the ultimate sources of their actions. My argument undercuts Tognazzini’s point.
control the relevant parts of her environment. (An intentionless force, such as Mele’s [2005] force field or some such, could also satisfy these two criteria because they could set Natasha up in this way and also locally control her environment. Of course, such an intentionless force would not know they were doing any of this. But, as I argued in §5.4, there is a sense of control that does not require intentionality.)

It can be argued that the covertly controlled Natasha is morally responsible by noting how similar this Natasha is to the merely manipulated Natasha. We have no reason to think that the merely manipulated Natasha is not morally responsible because she is no different to any other agent – viz. it is uncontroversial that the world plays some role in the causal aetiology of action. Psychologically and neurologically speaking, there is no relevant difference between the manipulated Natasha and any merely determined individual.

The covertly controlled Natasha initially might seem different to the merely manipulated Natasha, and so it might seem that we cannot generalise our judgement about the merely manipulated Natasha to the covertly controlled Natasha. The fact that the covertly controlled Natasha seems different or ‘monstrous’ (Russell 2010: 164) in comparison to normal individuals seems to be why we might judge that she is not morally responsible. But, as I argued in §5.7, while incompatibilists can make their case without explicitly endorsing the ‘no difference’ premise, they cannot deny that the ‘no difference’ premise is true, because this would be tantamount to conceding that there is some condition on moral responsibility that a covertly controlled individual has not met. So the covertly controlled Natasha is not different from the merely manipulated Natasha. Psychologically and neurologically speaking, there is no relevant difference between the manipulated Natasha and the covertly controlled Natasha. The motivational poster and the ego button have exactly the same effect on their respective brains and mental lives. The ego button, in effect, is a kind of internalised and unconscious motivational poster. Hence, we have no reason to think the covertly controlled Natasha is not morally responsible; indeed, the prima facie judgement that Natasha is not morally responsible
has been shown to be unreliable because an opposing judgement can easily be generated. Therefore the Ego Button is not a counter-example to all plausible current compatibilist conditions on moral responsibility.

As other unthreatening control cases share the same general structure as the Ego Button, this response can be applied to them too. In each of these cases, an individual is set up in some way – i.e. via design, conditioning, or programming – and then the controller employs some additional factor, such as knowledge of the future or local control of the environment, to secure control over the respective victim in these cases.\textsuperscript{144} Incompatibilists must accept that there are no relevant differences between such individuals and analogous merely determined individuals. Given that compatibilists hold that merely determined individuals are morally responsible, they can conclude that these covertly controlled individuals are also morally responsible. So, unthreatening control cases (as their name suggests) are unproblematic for compatibilism.

7.3 Hard vs. medium compatibilism

The sort of reply that I have proposed is similar to a response to control cases defended by McKenna (2008a, forthcoming) and Fischer (2011). McKenna classifies his reply as a ‘hard-line’ reply because it rejects premise 1 of MAS (that is, the manipulation argument schema), which I discussed in Chapter 5. In other words, his reply denies that covertly controlled individuals are not morally responsible. While there are similarities between these two responses, there are important dissimilarities between them too. As I see it, the ‘hard-line’ reply is a really a \textit{medium} compatibilist reply – that is, it is stronger than a soft compatibilist reply but weaker than a hard compatibilist one. After outlining the medium compatibilist reply, I will sketch a genuinely hard compatibilist reply that effectively extends the medium compatibilist reply. This move is important to set up my arguments in the following sections.

\textsuperscript{144} Examples of these cases are: Mele’s (2006) Designed Ernie, Pereboom’s (2001) Case 2, Kane’s (1996) \textit{Walden Two}-style case, Ginet’s (see Fischer 2000) \textit{Truman Show}-style case.
McKenna’s response to manipulation arguments in effect flips the argument on its head.

Recall that MAS goes as follows:

M1. (Judgement) An individual covertly controlled in manner X to perform action $A$ is not morally responsible for $A$-ing.

M2. (No difference) There are no differences relevant to moral responsibility between an individual who is covertly controlled in manner X to $A$ and an individual who is causally determined to $A$.

Therefore,

M3. If causal determinism is true, then no one is morally responsible for anything they do. In other words, compatibilism is false.

A control case is required to elicit the non-responsibility judgement to support M1, and, from the conjunction of M1 and M2, M3 is supposed to follow. To resist the generalisation of the non-responsibility judgement from a control case to a determination case McKenna’s rejects M1, and he does so because he believes a judgement which opposes, and therefore which undercuts, the non-responsibility judgement can be generated by reversing the order of the cases. So, the incompatibilist seeks to undermine compatibilism by starting with a control case and then generalising to a mere determination case; McKenna, however, runs it the other way. He starts with a mere determination case and then generalises to a control case. The result is that whatever judgement the determination case elicits can be generalised to the control case. Here is a formalisation of this move:

MC1. (Judgement) Causally determined individuals are morally responsible.
M2. (No difference) There are no differences relevant to moral responsibility between an individual who is covertly controlled in manner X to A and an individual who is causally determined to A.

Therefore,

MC3. An individual who is covertly controlled in manner X to A is morally responsible.

Clearly MC3 opposes M1; so it looks as if reversing the order of the argument undercuts the judgement which is required to get a manipulation argument off the ground.

Of course, I argued in §5.7 that incompatibilists can conclude that compatibilism is false if there is a control counter-example to all plausible current compatibilist conditions on moral responsibility. Such a control argument does not include a ‘no difference’ premise, so it might seem that the medium compatibilist reply will be ineffective against this argument. But, as I noted in §7.2, while a proponent of a control argument does not need to explicitly endorse a ‘no difference’ premise (such as M2), she cannot positively deny that there is no difference between the control and mere determination cases either. If she did, then she would be implicitly conceding that there is a condition on moral responsibility that the covertly controlled individual has not met, and that would undermine her argument against compatibilism.

McKenna, though, doesn’t endorse exactly this medium compatibilist argument. He doesn’t claim that a merely determined individual is morally responsible; rather, he claims that it is not clear that a merely determined individual is not morally responsible. So McKenna actually defends an argument of the following form:

MC1*. It is not clear that causally determined individuals are not morally responsible.
M2. (No difference) There are no differences relevant to moral responsibility between an individual who is covertly controlled in manner X to \( A \) and an individual who is causally determined to \( A \).

Therefore,

MC3*. It is not clear that an individual covertly controlled in manner X to \( A \) is not morally responsible.

This argument might not seem very favourable for the compatibilist. After all, it starts with premise which claims something that compatibilists presumably deny. But, according to McKenna (2008a: 155), his aim is not to present a positive thesis in favour of compatibilism; it is to defeat the incompatibilist’s argument. Call McKenna’s response – as per MC1*-M2-MC3* - a weak medium compatibilist reply, and call the previous one (MC1-M2-MC3) a strong medium compatibilist reply.\(^{145}\)

Why not go with a strong medium compatibilist reply? It seems that McKenna’s implicit rationale is that control cases might be presented to neutral inquirers. According to Pereboom (2008a: 162), neutral inquirers are participants in the debate who are undecided between compatibilism and incompatibilism, but who are open to both positions. Pereboom claims that we shouldn’t put much stock in the compatibilist’s judgements about these cases. According to Pereboom, the same applies to incompatibilist judgements; after all, incompatibilists presumably judge that any causally determined individual is not morally responsible just because she has been causally determined. Both compatibilists and incompatibilists are resolute in their judgements and therefore will not be easily swayed to agree with the opposing side’s judgement about a case. Since neutral inquirers judge that it is not clear that causally determined agents are not morally responsible, due to their neutrality, McKenna has perhaps framed his response to be sensitive to

\(^{145}\) Fischer’s (2011) medium compatibilist reply is a strong one.
this target audience. I think, then, that a medium compatibilist should not simply endorse a strong or a weak reply. Their official stance should be that the strength of the reply depends on the target audience of the argument. If the incompatibilist presents a control case to a compatibilist, then a strong reply is suitable. If, on the other hand, the incompatibilist presents a control case to a neutral inquirer, then a weak reply is suitable. Either way, the incompatibilist argument seems to be neutralised.

Medium compatibilism, however, as I argued in §6.3, falters in response to continuous manipulation cases. Consider Pereboom’s Case 1 – namely, the version where Plum1 only reasons egoistically when he is being covertly controlled by the neuroscientists. In this case, Plum1 is covertly controlled via sporadic continuous manipulation. That is, once the manipulators press their buttons, Plum1 is controlled moment to moment through the active manipulation period. Because of the extent of this manipulation, it seems clear even to compatibilists that Plum1 is not morally responsible. Medium compatibilists claim we can undermine that judgement by generalising our judgement about an analogous mere determination case to this control case. But Case 1 seems to elicit a non-responsibility judgement so strong that it cannot be neutralised via the reverse generalisation strategy. (Of course, I’ve argued in §6.4 that Case 1 is not, in fact, a counter-example because we misidentify the narrative self in Case 1.)

McKenna (2008a: 148-150), it seems, is sensitive to this worry. He in fact endorses a soft compatibilist response to this case. McKenna’s view is that we should endorse soft compatibilism as far as we can, and only after there are no more soft compatibilist replies available should we endorse the medium compatibilist response. But the soft compatibilist response to Case 1 that McKenna endorses is Fischer’s (2004) and Mele’s (2005), and both of

\[146\] Fischer’s soft compatibilist reply in this paper is only a reference back to a brief point from Fischer and Ravizza (1998). The response to Case 1 that Fischer develops in this paper is actually closer to a hard compatibilist reply. He argues that Plum1 is morally responsible, but not blameworthy due to the manipulation. On Fischer’s view, one can be morally responsible for \(A\)-ing without being praiseworthy or blameworthy for \(A\)-ing. While Fischer has a good reason (namely that we can be morally responsible for morally neutral actions too) for this view, his point can be
these required endorsing historicism, which I argued in §2.7 is untenable. Medium compatibilists might, then, endorse Demetriou’s (2010) soft compatibilist reply. But I argued in §6.3 that her reply is unsuccessful. Medium compatibilists could instead endorse my hard compatibilist reply to continuous manipulation, developed in §6.4, and endorse medium compatibilism in response to unthreatening control cases.

This raises a potential problem for a weak medium compatibilist reply (that is, one aimed at a neutral inquirer). While some neutral inquirers might happily generalise their neutral judgement about a merely causally determined individual to a covertly controlled individual and then conclude that the non-responsibility judgement about a covertly controlled individual is unreliable, some neutral inquirers might not. That is, it is plausible that some neutral inquirers will retrospectively amend their initial judgement about a normal deterministic agent even when faced with unthreatening control cases. Because of this, I think that medium compatibilist replies do not go far enough. While they might be enough to convince compatibilists that control cases do not pose a threat to their position, control cases might be enough to convince some neutral inquirers that compatibilism is theoretically costly. Compatibilists should explain away these apparent theoretical costs. If compatibilists don’t do this, then some neutral inquirers might be moved towards incompatibilism, and that would be bad for compatibilism.

In the remainder of this section, I’m going to sketch why I think natural compatibilists get the judgements they do about unthreatening control cases. Natural compatibilists are those compatibilists who find unthreateningly covertly controlled individuals to be morally responsible without explicitly considering the reverse generalisation strategy. The purpose of doing this is to set up my argument to sway the neutral inquirer towards compatibilism, which I present in §7.5.

As I’ve noted, I think that compatibilists must accept that certain covertly controlled individuals are morally responsible. After all, compatibilism is the thesis that moral responsibility
is compatible with causal determinism, and an implication of determinism is that our actions have causal sources in events external to us, and some of those external events might compose or constitute an agent. Some compatibilists might initially judge that an unthreateningly covertly controlled individual, such as Natasha, is not morally responsible. However, such compatibilists can render that judgement unreliable by endorsing medium compatibilism – that is, the reverse generalisation strategy.

Like I said above, natural compatibilists are those that judge unthreateningly covertly controlled individuals to be morally responsible without explicitly considering the reverse generalisation strategy. As far as such compatibilists are concerned, individuals like Natasha just are morally responsible, despite the fact they have been covertly controlled. But, just because natural compatibilists do not explicitly consider the reverse generalisation strategy, does not preclude them from implicitly considering it. I contend that natural compatibilists’ judgements stem from them internalising the reverse generalisation strategy. So, when a natural compatibilists considers (for example) the Ego Button case, she implicitly asks ‘is this case just like a case of mere determination?’, if it is, then she generalises her judgement about mere determination cases to the Ego Button case. I don’t think this process has to occur consciously; in fact, I think that it normally occurs unconsciously. Making the reverse generalisation strategy explicit is a way for natural compatibilists to turn non-natural compatibilists into natural compatibilists – i.e. by getting non-natural compatibilist to internalise the reverse generalisation strategy. I think that compatibilists can only benefit from making clear how they think about these cases. But, as I’ve said, I don’t think this strategy goes far enough.

On this picture, the ‘no difference’ premise has a role in the production of the natural compatibilist’s judgement about an unthreatening control cases. I propose that compatibilists take the medium compatibilist reply a step further. While medium compatibilists accept the ‘no difference’ premise for the sake of argument, compatibilists should instead argue that the ‘no difference’ premise is true. Recall that in §5.5 I argued that incompatibilists could endorse a
manipulation argument against compatibilism only if they modify the conclusion of that argument so that it reads:

M3*: If causal determinism is true, no one is morally responsible for anything they do because all causally determined individuals are covertly controlled by events external to them.

Some compatibilists might be tempted to reject M3* because they believe it is false that all causally determined individuals are covertly controlled by events external to them. However, I don’t think that will ultimately help defend compatibilism. I think compatibilists should instead accept the clause after the ‘because’ in M3* – that is, they should claim that all causally determined individuals are covertly controlled by events external to them – but still deny that M3* is true.Compatibilists, after all, have to accept that all causally determined individuals’ actions have sources in events external to them and that certain covertly controlled individuals are morally responsible, so it isn’t that much of an extension to this view to also claim that all causally determined individuals are covertly controlled. Compatibilists then have an argument for why there are no relevant differences between causally determined and unthreateningly covertly controlled individuals – viz. both are victims of unthreatening covert control. But since compatibilists hold that causally determined individuals (who satisfy the appropriate conditions, of course) are morally responsible, they also find that other victims of unthreatening of covert control, such as Natasha, are morally responsible too. Indeed, it might be that natural compatibilists already implicitly hold the view I am suggesting that all compatibilists endorse, and this might also explain why they judge that (for example) Natasha is morally responsible. On this view, Natasha is just one among many covertly controlled individuals, and so there is nothing obviously responsibility-undermining about the causal story of her action.
As I said earlier, some neutral inquirers might reject the non-responsibility judgement once they consider the reverse generalisation strategy, while neutral inquirers might not; what I’ve said here doesn’t add to that. The initial purpose of this move is just to make clear what the compatibilist’s commitments are, and how natural compatibilist’s think about unthreatening control cases; the eventual purpose will be to support an argument that should sway neutral inquirers towards compatibilism. I present this argument in §7.5. In the next section (§7.4), I argue that Todd’s moral standing manipulation argument neither shows that compatibilism is false nor increases its theoretical cost. My response to Todd allows me to explain some of the more nuanced features of compatibilism, which are often swept under the rug in the compatibility debate. Making these features prominent shows the benefits of compatibilism, and thereby decreases its apparent theoretical cost.

7.4 Covert control and moral standing

Patrick Todd (2012) has recently presented a new sort of manipulation argument. Unlike traditional manipulation arguments (i.e. straight or inference to the best explanation arguments, which I discussed in Chapter 5), Todd doesn’t try to claim that a ‘no difference’ premise is true; rather, he argues that an apparent implication of compatibilism is that a manipulator can erroneously blame her victim. For example, consider Designed Ernie*. In this case, Diana covertly controls Ernie* so that he murders Francis thirty years later. Ernie* satisfies all plausible current compatibilist conditions on moral responsibility when he acts. Compatibilists hold that Ernie* is morally responsible. Todd, however, doesn’t focus on whether or not Ernie* is not morally responsible; instead, he asks: ‘can Diana blame Ernie[*]? That is, having once determined for Ernie[*] to kill Jones, can Diana later blame him for doing so?’ (Todd 2012: 4).

Todd claims he finds it counter-intuitive that we blame Ernie* for murdering Francis, but that ‘it

---

147 Todd talks about Ernie rather than Ernie*, and he calls Ernie’s victim Jones rather than Francis. Nothing rides on these slight changes between our respective cases. Also, Todd talks about manipulation rather than covert control. His point applies, and indeed seems much stronger, when put in terms of covert control.

198
seems much more counterintuitive to suppose that Diana may blame Ernie[*] for what she
determined him to do’ (2012: 4). Let’s set aside the question of whether or not it is counter-
intuitive for us to blame Ernie* so that we can focus on the question of whether it is counter-
intuitive to suppose that Diana may blame Ernie*, and what implications this has for
compatibilism.

I agree with Todd that it is counter-intuitive to suppose that Diana may blame Ernie* for
killing Francis. But I disagree with Todd that the best explanation of this fact is that
incompatibilism is true. It seems to me that all the compatibilist simply has to say is that Diana
lacks the moral standing to blame Ernie*. An individual S might lack the moral standing to blame
another individual R because, for example, S is guilty of the same sorts of things as R, and so it
would be hypocritical of S to blame R. Or it might be because S and R are jointly responsible for
the same action or event, so it would be hypocritical of S to blame R (and vice versa) because S is
guilty of exactly the same crime as R. In a similar way, Diana is also responsible for Francis’s
death because she covertly controlled Ernie* to kill him; so it would be hypocritical of Diana to
blame Ernie* for killing Francis. In other words, Diana lacks the moral standing to blame Ernie*.

Todd does claim that the compatibilist best response to his argument is to claim that
Diana lacks the moral standing to blame Ernie*. However, at this point, Todd substitutes Diana
for God – an omnipotent, omniscient, and morally perfect being. He then argues that God
cannot lack the moral standing to blame anyone. I won’t rehearse Todd’s arguments because I
agree with Todd that God cannot lack moral standing in any way by virtue of being a morally
perfect being. After all, if God did lack moral standing, he wouldn’t be a morally perfect being. I
argue elsewhere that this substitution is crucial for Todd’s argument and effectively turns his
argument into a version of the logical problem of evil, and I argue that this only shows that a

148 Angela Smith (2007) argues that an individual’s moral standing affects whether or not she can blame an individual
morally responsible for something.
compatibilist free will response to the logical problem of evil fails. So here I will only consider Todd’s argument with Diana as the manipulator.

The view that manipulators cannot legitimately blame their victims has been proposed by Paul Russell (2010) as a response to control cases. While Russell doesn’t talk in terms of moral standing and instead talks in terms of manipulators not being able to take the ‘participant stance’, which expresses his sympathy with a Strawsonian account of moral responsibility, his point is not strictly speaking a Strawsonian one (though it is one to which Strawsonians are likely to be sympathetic). As Russell says, ‘[t]he strategy I … [defend] involves drawing a distinction between those who can and cannot legitimately hold an agent responsible in circumstances when the agent is being covertly controlled’ (2010: 168). Russell is claiming that there are conditions on who can and who cannot legitimately hold a particular individual morally responsible for her actions. This seems correct, and it seems that even a basic desert conception of moral responsibility should be sensitive to such conditions. I take Russell’s main point to be that:

The individual who covertly controls the agent is in no position to take up the participant stance towards this individual – no matter how “complex” or “robust” the capacities and qualities of the (created) agent may be. Having said this, similar constraints and limitations do not apply to other individuals who stand in a relevantly different relation to the covertly controlled agent (“Frankenstein”). Given that the

---

149 See ‘Manipulation, Moral Standing, and the Problem of Evil’ (manuscript) for more. In short, I argue that if God covertly controls Ernie* to kill Francis, then what we’ve got is really just an instance of the logical problem of evil. This is the problem of reconciling the fact God is supposed to be omnipotent, omniscient, and morally perfect with the existence of evil. Now if God covertly controls Ernie to kill Francis (which, by hypothesis, is a morally evil action), then it seems God cannot be morally perfect. I argue that Todd’s argument, at most, shows that the truth of compatibilism precludes the existence of God. But I argue that this is no reason to think that compatibilism is false because there is no conceptual link between an adequate free will response to the logical problem of evil and the true account of free will/moral responsibility. The only upshot is that compatibilists cannot be classical theists, and I think that would be a considered a benefit, rather than a cost, for most compatibilists.

150 I haven’t discussed Strawson’s (1962) view or Strawsonianism in this thesis due to space limitations. But this does not mean that I dismiss such a view. I think there are important lesson to learn from Strawsonianism. One of these lessons is that any plausible account of moral responsibility must include conditions which specify who is and who isn’t able to hold an individual morally responsible for her actions.
agent is not in any way impaired in his powers of rational self-control (i.e., he is not “abnormal” or “monstrous” in time-slice terms), and he is not covertly controlled by these other individuals, then the agent remains an appropriate target of “their” reactive attitudes. For these individuals, therefore, the participant stance is not ruled out or compromised simply on the ground that some other individual covertly controls him. (Russell 2010: 164)

So, when an individual S covertly controls an individual R to A, this does not imply that R is not morally responsible; rather, it only implies that S cannot legitimately hold R morally responsible for A-ing. Other individuals who did not covertly control R may still legitimately hold R responsible for A-ing if R satisfies all the compatibilist conditions on moral responsibility. After all, R is not relevantly different from other individuals, as she satisfies all the compatibilist conditions on moral responsibility (‘the [covertly controlled] agent is not in any way impaired by his powers of rational self-control’). On Russell’s view, S cannot take the participant stance with respect to R because ‘there is insufficient causal distance between [a manipulator and her victim]. Moral communication and responsiveness presuppose that agents are not related in this way.’ (2010: 159). I think that what grounds the fact that moral communication and responsiveness presuppose that agents are not related in this way is that legitimate moral communication and responsiveness can only be had if agents have the moral standing to hold each other responsible for their action. So, we might say S cannot legitimately take the participant stance to R with respect to A because S lacks the moral standing to hold R morally responsible for A-ing.

It should be clear how Russell would respond to Todd’s point: he would say that Diana cannot take the participant stance towards Ernie*. And what I think underlies Russell’s point is that Diana lacks the moral standing to blame Ernie* for his actions because she too is morally responsible for them. Incompatibilists might object here. They might claim that if Diana is
morally responsible for Ernie*’s actions, then Ernie* cannot be morally responsible. They might insist that only one individual may be morally responsible for a particular action or event. But compatibilists should deny this. As Frankfurt says:

There is no paradox in the supposition that a [manipulator] might create a morally free agent. It might be reasonable, to be sure, to hold the [manipulator] morally responsible for what his free subject does, at least insofar as he can be fairly held responsible for anticipating the subject’s actions. This does not imply, however, that full moral responsibility for those actions may not also be ascribable to the subject. It is quite possible for more than one person to bear full moral responsibility for the same event or action. (1975/1988: 54)

Incompatibilists, such as Todd, might deny Frankfurt’s claim. They might insist, as noted, that only one individual may be morally responsible for a particular action or event. However, it is only possibly true that one individual may be morally responsible for a particular action or event if incompatibilism is true. For example, according to source incompatibilists, an individual is morally responsible only if she is the ultimate source of her actions. Being the ultimate source of an action seems to be something only one individual can be – that is, it seems that two individuals cannot both be the ultimate source of a particular action or event; hence, such a view seems to entail that only one individual may be morally responsible for any particular action or event. But the truth of (source) incompatibilism hasn’t been established yet, so incompatibilists cannot help themselves to such a thesis on pain of begging the question against the compatibilist. Incompatibilists would have to argue for this view without presupposing that compatibilism is false, and I can’t see how this can be done.

---

151 This point also requires rejecting the possibility of joint actions. Since it would unnecessarily complicate matters, I set this point aside here.
While I make no stand on the plausibility of a Strawsonian account of moral responsibility, I think that all accounts of moral responsibility should embrace the insight that there are conditions on who can and who can’t legitimate praise or blame an individual for an action she is morally responsible for. Pereboom (2014) has recently made two points in response to Russell which seem to implicitly reject these insights. Responding to Pereboom’s point will thereby emphasise the plausibility of (hard) compatibilism.

Pereboom first claims that:

… even if it is inappropriate for [a manipulator] to express indignation toward [her victim] or to punish him for retributive reasons, presumably it will be legitimate for [the manipulator] to encourage or to causally determine an intermediary to do so. The intermediary did not cause [the covertly controlled individual] to act badly, so she is in a position to hold him morally responsible, and thus [the manipulator’s] causing the intermediary to hold him responsible would amount to [the manipulator] causing her to perform a permissible (or even a required) action.152 (2014: 97-98)

Pereboom then claims that it would be counter-intuitive to accept that the intermediary of the manipulator could legitimately blame the victim because the intermediary is caused to do so by the manipulator. However, it is unclear to me that this is counter-intuitive. Russell’s point, which Pereboom seems to acknowledge, is that the manipulator, by covertly controlling her victim, makes it inappropriate for her to hold the victim morally responsible (with respect to the actions she covertly controls the victim to perform). The fact that the manipulator can covertly control (that is, ‘encourage’) other individuals to blame the original covertly controlled individual actually emphasises Russell’s point: by virtue of being covertly controlled, the ‘intermediary’ and the

152 Pereboom is actually talking about God and a Pharaoh in this quotation. Pereboom takes God to be playing the role of a manipulator and the Pharaoh to be his victim. My response to this example would be the same as my response to Todd’s argument where he invokes God as the covert controller of Ernie*.
covertly controlled individual are not relevantly different from each other; therefore they can 
legitimately praise or blame each other. It is no problem for the hard compatibilist, then, for the 
manipulator to covertly control an intermediary to blame another individual whom the 
manipulator has covertly controlled to perform some action.\textsuperscript{153}

The reason why it might \textit{seem} illegitimate for the manipulator to encourage others to 
blame the victim is that it seems to be a way for the manipulator to get themselves off the hook 
for the victim’s action in the eyes of other people. But the fact that the manipulator can 
legitimately encourage others to blame the victim does not get her off the hook. As the covert 
controller of the victim, the manipulator is also morally responsible for the victim’s actions. The 
way Pereboom describes his case is such that \textit{only} the victim gets the blame, and that seems to 
imply the manipulator isn’t then blameworthy. As we’ve already seen, two individuals may be 
morally responsible for the same action or event, so it remains that the manipulator is 
blameworthy and, thus, can be justifiably be blamed. Hence, the fact that the victim can 
legitimately be blamed by certain appropriate parties in no way undermines the fact the 
manipulator can also be legitimately blamed.

Pereboom’s second point is that, ‘it’s also plausible that this putative inappropriateness 
on [the manipulator’s] part would be prima facie and limited. On the basic desert conception, 
holding morally responsible is not merely a privilege, but rather a moral good, and one that can 
either be realized or not’ (2014: 98). Pereboom’s thought seems to be that it is irrelevant what 
role the manipulator had in the production of an action. If an individual is morally responsible 
for an action, then she deserves praise or blame for that action. And if an individual deserves 
praise or blame, \textit{somebody} – it doesn’t matter who – ought to be doing the praising or blaming.

Because there is a moral good to be had in praising or blaming a morally responsible individual,

\textsuperscript{153} This assumes that the intermediary satisfies the compatibilist conditions on moral responsibility, and hence is a 
morally responsible agent. If the intermediary were not a morally responsible agent, we could simply argue that the 
intermediary cannot blame the victim because she is an extension (or some such) of the manipulator.
if the only person available to do the blaming might initially be inappropriate to do so, that initial inappropriateness can be outweighed by the good that would be realised in blaming being done.

Pereboom’s proposed view, however, is implausible. There are plenty of cases in which it would be inappropriate for just anyone to blame someone for a wrongdoing. For example, suppose Barry is mean to his partner, Larry, at a dinner party. It turns out that the dinner party is full of couples where one partner is often mean to the other partner. While it seems appropriate for Larry to blame Barry for his meanness, it seems inappropriate for all the other mean partners to blame Barry; it is inappropriate because the mean partners lack the moral standing to blame Barry for his action – that is to say, it would be hypocritical of the mean partners to blame Barry for the kind of thing that they also do. As I’ve claimed, it seems that any account of moral responsibility must make room for features like this. Even if we accept a basic desert conception of moral responsibility and we accept that it is a moral good for a blameworthy individual to be blamed, it does not follow that this moral good can be realised by anyone blaming the blameworthy individual. Hence, it is not always appropriate for just anyone to blame an individual for actions she is blameworthy for. Accounts of moral responsibility, then, must include ‘community’ conditions – that is, norms governing when it is permissible to praise or blame another individual for her actions. So, while compatibilists claim that a covertly controlled individual is morally responsible and therefore can be legitimately praised or blamed for her actions, this does not imply that anyone can legitimately praise or blame the covertly controlled individual. Even if there is a moral good to be had from praising or blaming a covertly controlled individual, this moral good can only be realised by certain parties.

Pereboom then makes a further claim to support his second point. He writes:

If the good of holding [a covertly controlled agent] morally responsible for say, twenty seriously bad actions would have to be foregone unless [her manipulator] were to do so, it would seem that at some point the good of [the manipulator] taking
action would outweigh the inappropriateness of [the manipulator] playing this role.

(Pereboom 2014: 98)

Pereboom point seems to be that it is possible that it would become appropriate – indeed, a positive good – for the manipulator to blame her victim, if it meant that otherwise the victim was not blamed by anyone. But there’s no reason why a compatibilist should accept this because it is not entailed by a basic desert conception of moral responsibility. As I argued above, it seems that the ‘moral good’ of blaming a blameworthy individual can only be realised by an appropriate party doing the blaming. For example, the moral good of blaming Barry can only be realised by Larry (or one of the non-mean partners); it cannot be realised by one of the mean partners.

Pereboom might claim that if all the non-mean partners suddenly died (and everyone else in existence too), then it would become appropriate for a mean partner to blame Barry for being mean. But I can’t see how this is entailed by the basic desert thesis that is supposed to be at issue in the compatibility debate. If it’s correct that the moral good of blaming an individual can only be realised by appropriate individuals doing the blaming, then it is unclear how it can become appropriate for other previously inappropriate individuals to the blaming.

This highlights something which I think Pereboom illicitly trades on in his argument against compatibilism. He claims that the sense of moral responsibility at issue is a basic desert one. As I noted in §1.4, I have accepted this in this thesis. But this doesn’t mean that compatibilists need to accept everything Pereboom claims (implicitly or explicitly) is implied by a basic desert thesis. According to Pereboom, desert is ‘basic’ because the justification for praising and blaming stems just from an individual being morally responsible for a particular action, and not from consequentialist or contractualist considerations. But compatibilists are open to endorse their own account of what of the content of ‘praising’ and ‘blaming’ is. For example, a compatibilist need not accept that individual who steals someone else must be openly hated and resented for stealing. This might be justified, depending on the details of the crime and the
criminal, but it might in fact be appropriate to blame the thief in some other way. For instance, when we are wronged we don’t always overtly blame the culprit, but we can usually still *privately* blame the culprit (as I discussed in §1.2). Of course, it might be tenuous to call these private responses ‘blame’. But compatibilists can argue that a morally responsible individual deserves *some overt response* just because she has performed an action. While we might privately blame an individual by (for example) resenting them, we might overtly respond to them in a more neutral ways. For example, we could tell them that what they have done is morally wrong, that they shouldn’t be doing such things, and that they should change their ways. When we respond to a morally responsible individual in such ways, we do not express any blaming emotions (though we might be feeling them), but we still express a blaming attitude – that is, we make clear that that we consider the individual to be responsible for a morally wrong action.

I think that the appropriate content of our private and overt responses will depend on further factors such as, for example, what theory of punishment or what moral theory is true. As long as the justification for the response to an action is basic, compatibilists are still endorsing the basic desert account of moral responsibility that Pereboom claims is at issue in the compatibility debate. Such a basic desert view does not, however, come with all the implications which Pereboom (or any incompatibilist) builds into *his* basic desert thesis. Fully developing this aspect of compatibilism is something I lack the space to do here. But notice that compatibilism is not forced into overt retributive responses to morally wrong actions. Compatibilists can still be constructive or restorative in their responses to morally wrong actions while still claiming that such responses are basically deserved.

As we’ve seen, incompatibilists (such as Pereboom and Todd) have tried to make compatibilism seem implausible by arguing that it has costly implications. But, as I argued, compatibilism – or at least *hard compatibilism*, the view that covertly controlled individuals (including all causally determined individuals) may be morally responsible for their actions – does not have such implications.
7.5 The collapse of source incompatibilism

My final argument for hard compatibilism is a destructive one. I will argue that, just as historicism collapses into source incompatibilism, source incompatibilism collapses into impossibilism – the thesis that free will and moral responsibility are impossible. I then assess the plausibility of a ‘no covert control’ condition on moral responsibility. I argue that such a condition is patently ad hoc in a dialectic context which includes the impossibilist. Thus, control cases leave only two principled positions: hard compatibilism and impossibilism. I conclude that we, including neutral inquirers, should prefer hard compatibilism to impossibilism.

Source incompatibilists, like historicists, use control cases to motivate their view. Source incompatibilist – like historicists – therefore implicitly accept that control cases are an adequate test of the conditions on moral responsibility. Because they accept this, if there is a control case which is a counter-example to the source incompatibilist conditions on moral responsibility, then source incompatibilists must accept that the source incompatibilist conditions are insufficient for moral responsibility; otherwise source incompatibilism is methodologically inconsistent in the same way I argued in §2.7 that historicism is. Since source incompatibilists motivate their position by an appeal to control cases, they cannot bite the bullet and claim that such cases are not counter-examples. Hence, a successful control counter-example to source incompatibilism will show that source incompatibilism is false.

The problem that control cases are supposed to highlight for compatibilism is that a causally determined individual’s actions have causal sources in events external to her. But source incompatibilists cannot claim that moral responsibility requires that the production of an action is completely unaffected by events external to the individual. For an individual to be completely

154 The ‘direct argument’ – a variant of van Inwagen’s consequence or indirect argument that argues that determinism rules out moral responsibility without mentioning alternative possibilities – may also been used to prima facie motivate source incompatibilism, as I noted in n.21. As I mentioned there, McKenna (2008c) argues that the success of the direct argument is parasitic on the success of a control case/manipulation argument. Given this, control cases are the only definitive motivation for source incompatibilism.
unaffected by external events, an individual would have to be \textit{causa sui} – that is, self-creating – and that seems to be impossible for beings like us (cf. G. Strawson 1986, 1994). Hence all source incompatibilists – whether they endorse event-causalism or agent-causalism – must accept the role of the external world in the production of action. But this means that there is nothing \textit{a priori} that rules out the possibility of covertly controlling a source incompatibilist (that is, source libertarian) individual. To achieve this, I contend that a manipulator need only do what she does to covertly control a deterministic individual – namely control factors such individuals have no control over, such as their initial character and their environment.

My strategy in the following will be to propose a new control case that initially purports to be a counter-example to all plausible current compatibilist conditions on moral responsibility. I’ll then stipulate that the victim in this case – Alana – satisfies the source incompatibilist conditions on moral responsibility, and I’ll argue that Alana is still covertly controlled in these cases; therefore if we judge that a deterministic individual is not morally responsible when she is covertly controlled, then we should judge that a source libertarian individual is not morally responsible when she is covertly controlled. Here’s the case:

\textit{Alana and the Aliens}. Humphrey, an alien with extensive powers, plans to conquer Earth but he is worried that Clive will interfere with his plans. Humphrey decides to stop (i.e. kill) Clive, but elects to use a human agent to do his dirty work. To do this, he modifies a zygote that has just been conceived. This zygote will now develop into a nasty agent called Alana, as her zygote has been modified to have a genetic profile that is conducive to performing heinous actions later in life. To ensure that Alana will kill Clive, Humphrey controls Alana’s entire environment. He ensures that all her experiences are conducive to her killing Clive. When the time comes, Alana kills Clive, and Humphrey is careful to ensure that Alana satisfies all plausible current compatibilist conditions on moral responsibility when she acts.
As described, this case purports to be a counter-example to all plausible current compatibilist conditions on moral responsibility. Given that Alana has been covertly controlled while satisfying all plausible current compatibilist conditions, it seems that all those conditions are insufficient for moral responsibility, and hence compatibilism is false. Hard compatibilists, of course, will deny this. Hard compatibilists will argue that Alana is no different from any merely determined individual, so the non-responsibility judgement (in those that get it) is unreliable. (Of course, she is a very unlucky individual. Some of us are lucky enough not to encounter environments which are conducive to us becoming killers. But the fact that Alana is unlucky doesn’t undermine her moral responsibility for her actions; it might affect what response that action deserves, though.) But, for the sake of argument, let’s assume that such cases show that compatibilism is false. My point here is to show that source incompatibilism collapses into impossibilism, so this is a harmless assumption in this context. Source incompatibilists implicitly claim the non-responsibility judgement about agents like Alana is reliable, and cases involving agents like Alana are therefore claimed to motivate source incompatibilism. Underlying this motivation is the claim that being covertly controlled is necessarily responsibility-undermining; it is because of this that individuals like Alana seem to lack moral responsibility.

To show that source incompatibilism is false, we must therefore only show that an individual can be covertly controlled to A whilst satisfying the source incompatibilist conditions on moral responsibility. There are two main forms of source incompatibilism: an event-causal view and an agent-causal view. According to the former, the causal story of free action must feature indeterminism at some point. The event-causal view is pretty much a compatibilist view but with the addition that for an action to be a free action there must be some indeterminism at some point in the individual’s deliberative history that is relevant to the act in question. According to the agent-causal view, a free action must be caused by the individual fundamentally as a substance, and not by the individual’s states. In what follows I will focus on showing that an
adapted version of Alana and the Aliens is a counter-example to an agent-causal source incompatibilist view. This will *a fortiori* show that it is also (suitably adapted) a counter-example to an event-causal view.\(^{155}\)

What unites source incompatibilist views, as I noted in §1.1, is the acceptance that robust alternative possibilities are not necessary for moral responsibility.\(^{156}\) But it won’t matter in what follows what kind of alternative possibilities we think Alana has. I will argue that whatever alternatives she has, she is still plausibly covertly controlled by Humphrey, and hence isn’t morally responsible for killing Clive.\(^{157}\)

Source incompatibilists might first claim that because we have assumed that determinism is false, we are not entitled to assume that Alana will develop into a nasty person. It is, of course, true that if we assume that determinism is false, we cannot say that Alana will definitely develop into nasty person. But this doesn’t mean Humphrey does not covertly control Alana no matter what she does, or so I shall now argue.

Let’s now stipulate that Alana is an agent-causal agent, so her action of killing Clive is caused by her fundamentally as a substance.\(^{158}\) According to the agent-causalists, Alana is now morally responsible for killing Clive. If this is true, then source incompatibilists must first explain why Alana has not been covertly controlled by Clive. After all, as I claimed above, source incompatibilists cannot accept that covertly controlled individuals are morally responsible without undercutting the motivation for their own view.

\(^{155}\) My reason for doing so is that it is somewhat uncontroversial that there are control counter-examples to event-causal source incompatibilist views (e.g. Pereboom 2002; Mele 2005; King 2013)

\(^{156}\) Kane (1996) calls himself a source incompatibilist, yet he claims that robust alternative possibilities are a necessary condition on being morally responsible. Shabo (2010) call this sort of source incompatibilism (somewhat pejoratively) ‘compromising’. The sort of source incompatibilism I am discussing is what Shabo calls ‘uncompromising’.

\(^{157}\) We could also imagine that Humphrey acts as a ‘counterfactual intervener’ which, as I discussed in Chapter 1, is the agent in a Frankfurt case whose presences stops an individual from having robust alternatives when they act.

\(^{158}\) As with apparent counter-examples to compatibilism, I will assume that Alana satisfies all the uncontroversial conditions on moral responsibility when she kills Clive. Thus, while being an agent-cause is only supposed to be a necessary condition on moral responsibility, we can treat it as a sufficient condition in this context.
There are two reasons why such an explanation does not seem available. First, there is nothing inherent in being an agent-cause which means that an individual cannot be covertly controlled. Ned Markosian (1999, 2012), for instance, has argued that deterministic individuals can also be agent-causal agents. To be sure, there is nothing obvious about why being able to cause one’s actions fundamentally as a substance is incompatible with the truth of causal determinism. If Alana were an agent-causal agent and causal determinism were true – something that Markosian argues to be a conceptual possibility – would she avoid being covertly controlled by Humphrey? It seems not. Regardless of how Alana causes her actions, Humphrey is still able to control her to do his bidding.

Second, even if Alana is an indeterministic agent-causal agent, she must still act for reasons. While she, by hypothesis, has control over which reasons she acts from, she cannot control what her reasons for action are. Suppose that in her situation with Clive, Alana has a reason to kill Clive (i.e. she hates him and wants him to die) and a reason to go and buy some ice cream instead (i.e. she loves ice cream and is currently quite hungry). And suppose Humphrey has locally controlled her environment to ensure that she has both these reasons to act. While he would prefer that Alana kill Clive, he’s also set things up so that she might just buy some ice cream instead. Humphrey, as we’d expect, has also set things up so that the probability of Alana choosing to buy ice cream is really low – let’s say 5%. To do this, Humphrey ensured that Alana was brought-up by nasty parents (with a love for ice cream), he made sure she didn’t make friends in school and ultimately become reclusive, and, among other things, he’s controlled things such that Alana only has experiences that teach her to value heinous and nasty behaviour, such as murdering those we do not like. While she is not directly being ‘given’ her reasons for action, Humphrey is clearly pulling the strings behind the scene to ensure that she has such reasons for action. If it ever looks like Alana is veering away from being a nasty person, Humphrey ramps up his local control of her environment. As a result, it is overwhelmingly likely that Alana will kill Clive.
Source incompatibilists could insist that Humphrey does not control Alana because she can do otherwise than kill Clive. In response, we might first stipulate that if Alana does do otherwise than kill Clive initially, that Humphrey will engineer things so that Alana gets another chance to kill him. And each time he engineers another chance to do so, he increases the odds that Alana will murder Clive. It looks as if eventually Alana will kill Clive. It seems that given the overwhelming odds that Alana will kill Clive and Humphrey’s role in stacking those odds, that Alana has been covertly controlled to kill Clive in the possible worlds in which she does kill him.

Even if there is some possible world where Alana does not kill Clive, remember that control cases are used to argue that indeterminism (whether in conjunction with agent-causalism or not) is a necessary condition on moral responsibility; more specifically, it is claimed to be (at least) necessary for the control or freedom (i.e. free will) required to be morally responsible for one’s actions. This means that being in control or controlling something is compatible with indeterminism – that is, with more than one physically possible outcome. If control is compatible with indeterminism, then it seems that Humphrey can still control Alana even if there is a chance that Humphrey cannot guarantee (i.e. causally determine) how Alana will act. Indeed, given that Humphrey has controlled which reasons for action she has throughout her life, it seems he controls her whatever she decides to do.\[159\]

Put it this way. Alana is like a puppet with a battery. Humphrey – the puppet master – has no control over the battery. It activates only under Alana’s ‘control’. But Humphrey controls everything else about Alana. It’s just that Alana needs to ‘activate’ at the right times. Humphrey is a good puppet master who knows his subject well, so there is a high probability that Alana will do what he would prefer her to do. Even if she doesn’t do that, Humphrey doesn’t mind. He’s

\[159\] Alternatively, we might think of control over other people as being an actual-sequence notion. So, as long as Alana does what Humphrey wants her to do in the actual sequence of events and Humphrey has a major causal role in Alana doing what he wants (e.g. by locally controlling her environment and creating her such that she inclined to perform nasty actions), then Humphrey controls Alana in the actual sequence. On this view, it doesn’t matter what Alana does in other possible worlds. She might not be controlled by Humphrey in those worlds, but in the actual world she is, and that’s what matters.
set her up so that the only options available to her are ones that he has ensured she has. And Humphrey will control things so that Alana is repeatedly put into situations where the odds increase that she will do what Humphrey really wants her to do (i.e. kill Clive). Alana has a power, of course, but Humphrey’s still pulling her strings; he’s still controlling her.

Of course, source incompatibilists might claim that there is a difference between an individual’s control over herself and control over others. The source incompatibilist could claim that while the former requires indeterminism, the latter requires determinism. In other words, we only control other people if we causally determine them to act. But that seems somewhat ad hoc. What reason, other than that it gets the source incompatibilist the result they want, could be given to support this difference? There doesn’t seem to be a principled reason why we should differentiate between the conditions required for controlling ourselves and those required for controlling others. Hence, given that Humphrey supplies all Alana’s reason to act, however she acts will be under Humphrey’s direction. Just as we sometimes do not act in the way we’d really prefer to act, a covert controller sometimes cannot get his victim to act how he really wants her to act, but that doesn’t mean he hasn’t exercised control over how she acts. Therefore there’s no good reason to discount the claim that Humphrey covertly controls Alana; so we should think that Humphrey does covertly control Alana. But if Humphrey covertly controls Alana, then Alana is not morally responsible for killing Clive. And if Alana is not morally responsible despite satisfying the source incompatibilist conditions on moral responsibility, then source incompatibilism is false.

It makes no difference that Alana was covertly controlled by an agent. We could easily replace Humphrey with an intentionless force without affecting the case. So, for example, we could imagine that an alien plant arrived on Earth via a meteorite. This plant then, as it does, modified a zygote in one of the local populous so that a particular individual – Alana – would emerge. The plant then – using its extensive powers – controlled Alana’s environment and thereby (in conjunction with modifying her zygote) covertly controlling her (completely
unintentionally, of course) to kill Clive. Of course, there remains a small chance that Alana
doesn’t kill Clive, but that does not undermine the alien plant’s control over Alana; it is in fact
consitutive of it in this context. Indeed, source incompatibilists couldn’t take this line of
response any way. If one argues that there is a relevant difference between agents and
intentionless forces, then this undercuts source incompatibilist’s argument for their own position
– namely the control/manipulation argument. Thus, source incompatibilists have implicitly
accepted that blind forces can covertly control individuals just as well as agents can.

The fact that Humphrey can be replaced with an intentionless force allows me to extend
my argument from §7.2. There I argued that the hard compatibilist should accept that all causally
determined individuals are covertly controlled by (at least) events external to them. Here I argue
that hard compatibilists should also claim if causally determined individuals are covertly
controlled by events external to them, then all individuals (well, at least human persons) are
covertly controlled by external events, including agent-causal source libertarian agents (and a
fortiori all other source libertarian individuals). The only difference between them is that the sort
of control differs – that is, causally determined individuals are deterministically controlled, while
agent-causal individuals are indeterministically controlled. Either way, both sets of individuals are
covertly controlled. Hence, if control cases undermine compatibilism, then they also undermine
source incompatibilism.

As source incompatibilist motivate their position by an appeal to control cases, they
cannot simply bite the bullet when faced with one which is problematic for their position. They
rather need to appeal to some principled difference between these control cases, but none seems
available. What control cases seem to show is that moral responsibility requires that no external
events influence our behaviour. Not only must we control which reasons we act for, but we must
be able to control what reasons are available to us. Only this way can be avoid being covertly
controlled by other agents or events external to us. What kind of condition must an individual
satisfy to guarantee not being covertly controlled? Well, it seems that the only available option is
that an individual be self-creating. But self-creation is impossible for creatures like us, so moral
responsibility is impossible. Hence, source incompatibilism collapses into impossibilism. The
only support for source incompatibilism – control cases – in fact supports impossibilism.

What about a ‘no covert control’ condition? If source incompatibilists could endorse this,
then they could resist the collapse into impossibilism. According to this condition, an
individual S is not morally responsible for an action \( A \) if S stands in a ‘covert control’ relation to
an individual R (or some intentionless force). To avoid the collapse into impossibilism, source
incompatibilists could perhaps endorse such a condition.

But source incompatibilists cannot endorse such a condition. First, if source
incompatibilists could endorse such a condition, then there’s no reason why compatibilists
couldn’t either. Source incompatibilists therefore cannot endorse this condition because it
provides compatibilists with a response to the control cases which are supposed to motivate
source incompatibilism in the first place. Hence, source incompatibilists would undermine the
motivation for their view by endorsing such a condition.

Second, while some compatibilists (e.g. Haji and Cuypers 2001; Mele 1995, 2006) have
proposed conditions similar to a ‘no covert control’ condition, such a condition is infelicitous in
this dialectic context because it implicitly dismisses impossibilism as a viable position. According
to the impossibilist, free will and moral responsibility require that we be self-creating. If we were
self-creating, then we couldn’t be covertly controlled. Thus, a ‘no covert control’ condition is
patently ad hoc because it is only posited to rule out control cases which could be used to
motivate impossibilism, and thereby get moral responsibility possibilities the result they want.
Given this, compatibilists and source incompatibilists cannot endorse such a condition.

\[160\] Haji and Cuypers (2001), and Mele (1995, 2006), endorse something like this condition, though they actually call
it a ‘no bypassing’ condition. This condition is motivated by cases like Brainwashed Beth. However, as I argued in
§2.7, historicists cannot justify different responses to Brainwashed Beth and Designed Ernie*. So it seems that a ‘no
bypassing’ condition collapses into a ‘no covert control’ condition. Lycan (1987/1995) explicitly suggests such a
condition, as I discussed in §5.3. Haji and Cuypers (2001) also argue that agent-cause individuals can be victims of
responsibility-underminating covert control (or ‘manipulation’ as they call it).
As we’ve seen, there are control cases that are counter-examples to source incompatibilism. The source incompatibilist lacks any principled reason for resisting these cases whilst using other control cases to motivate her position. It’s no good appealing to a ‘no covert control’ condition either. Not only could compatibilist’s use this, but it’s also ad hoc. Hence, there’s no stopping source incompatibilism’s collapse into impossibilism.

Only hard compatibilism is able to avoid the descent into impossibilism because only hard compatibilists claim that unthreatening control cases are not adequate tests of the conditions on moral responsibility. Hard compatibilists reach this conclusion because they accept that all individuals – whether they are causally determined or not – are covertly controlled by events external to them. Given this, any creepiness that unthreatening control might have produced soon dissipates because it is something all individuals (at least ones like us) are subject to. Consequently, I contend that the non-responsibility judgement should start to lose its pull – at least from a neutral inquirer’s perspective, though perhaps not from an incompatibilist’s.

Neutral inquirers now have a good reason to agree that covertly controlled individuals are morally responsible – namely that all individuals like us are covertly controlled in some way.

This sort of argument will likely only appeal to moral responsibility possibilists. I have presented no argument against those who are initially inclined to believe that moral responsibility is impossible. What I have shown is that if we believe that moral responsibility is possible, hard compatibilism is the only viable option. Given this, there is genuinely no theoretical cost that comes with hard compatibilism because it is the only option available for moral responsibility possibilists. If impossibilism is a genuine option – which I think it is – then we must weigh up which is preferable out of these two views. How might we choose between hard compatibilism and impossibilism? Well, we might weigh up the costs and benefits of each view, and then see which comes out on top. It seems clear to me that hard compatibilism, as it maintains our
practice of moral responsibility, wins such a contest. Hence, hard compatibilism is preferable to impossibilism.\textsuperscript{161}

7.6 Conclusion

In this and the previous chapter, I have argued that control cases are not counter-examples to all plausible current compatibilist conditions on moral responsibility, and therefore do not show that compatibilism is false. In response to continuous manipulation cases, I argued that continuous manipulation in fact creates a new agent. This new agent may then be morally responsible for her actions. In response to sporadic continuous manipulation, I argued that either (i) a new narrative self is created or (ii) no new narrative self is created, and so we can run the reverse generalisation strategy/medium compatibilist reply against it. In response to unthreatening control cases, I had three interlinked responses. First, I endorsed the medium compatibilist reply. This reply shows that the non-responsibility judgement elicited by unthreatening control cases is unreliable because an opposing judgement can easily be generated. I then developed this response by claiming that hard compatibilists ought to claim that all causally determined individuals are covertly controlled. Second, I endorsed Russell’s selective hard compatibilism. He argues that manipulators cannot blame their victims. I then defended this view from two objections of Pereboom’s (2014). Third, I argued that hard compatibilism is preferable to source incompatibilism because source incompatibilism collapses into impossibilism, and I argue that hard compatibilism is preferable to impossibilism because it is better to preserve our practice of moral responsibility than to jettison it. Hence, if we believe moral responsibility is possible, we must endorse hard compatibilism.

\textsuperscript{161} There might some who react to my argument against source incompatibilism by simply accepting that impossibilism is true (and, thus, implicitly reject the sort of cost-benefit analysis I have appealed to). Those inclined to be anti-compatibilist might be moved in this way. I think the apparent plausibility of impossibilism rests on an implicit assumption of \textit{free will realism} – that is, (roughly) the view that free will is natural kind. Once free will realism is exposed, though, there is no good reason to endorse it; hence we should endorse free will \textit{anti-realism}. Furthermore, I think that compatibilists are actually implicitly committed to free will anti-realism, and this helps to further explain why (natural) compatibilists have the judgements they do about these sorts of cases. Developing this view, however, is a matter for future work.
The position I have defended is a hard compatibilist one. While I accept that certain covertly controlled individuals are morally responsible – that is, ones which satisfy the synchronic ownership conditions of my structural-narrative view – I have shown that there is nothing unprincipled about this. Critics of hard compatibilism often like to paint this position as a mere bullet-biting response. And while this position does bite the bullet in a way, it does so by explaining why covert control does not necessarily undermine an individual’s moral responsibility, thereby making the bullet much more palatable.

The general lesson of hard compatibilism is that while covert control might have some lingering creepiness, this creepiness is an unavoidable creepiness if we believe that moral responsibility is possible. Given that this apparent creepiness is something that all views about moral responsibility must deal with and that no matter whether determinism or indeterminism is true or whether we are agent-causes or not, there is a sense in which all human persons are covertly controlled by events external to them; hence, because it is something we are all subject to, the creepiness of covert control no longer seems that creepy. Once we accept this, we should then see that covert control – in the nonconstraining form used in unthreatening control cases – is not a genuine threat to our free will or moral responsibility for our actions. Hence control or manipulation arguments do not undermine compatibilism.

***

In this thesis, I have defended compatibilism from all control/manipulation arguments. In Part One (Chapters 2-4), I argued that compatibilists must endorse a structural or non-historical account of moral ownership (Chapter 2), and I defended my own version, the structural-narrative view (Chapters 3 and 4). In Chapter 5, I clarified the structure of manipulation arguments and I argued that the real threat to compatibilism comes from control cases and their associated argument. In these final two chapters, I proposed a hard compatibilist response to all the remaining control cases. I have shown that this view much more principled position than its
critics often paint it to be. While there is certainly more work to be done to develop the view I have defended here, my defence is more than enough to show that control/manipulation arguments do not undermine compatibilism nor do they make it theoretically costly.
Bibliography

Alvarez, Maria (2009). ‘Actions, thought-experiments and the ‘principle of alternate possibilities”.


Cambridge.


------ (2013). ‘Historicism, Non-historicism, or a Mix?’ *Journal of Ethics*, 17, 3: 185-204.


Locke, John (1690) An Essay Concerning Human Understanding.


229


