Philos Stud (2014) 170:317–334
DOI 10.1007/s11098-013-0220-9

Compatibilism and personal identity

Benjamin Matheson

Published online: 17 October 2013

Abstract Compatibilists disagree over whether there are historical conditions on moral responsibility. Historicists claim there are, whilst structuralists deny this. Historicists motivate their position by claiming to avoid the counter-intuitive implications of structuralism. I do two things in this paper. First, I argue that historicism has just as counter-intuitive implications as structuralism when faced with thought experiments inspired by those found in the personal identity literature. Hence, historicism is not automatically preferable to structuralism. Second, I argue that structuralism is much more plausible once we accept that personal identity is irrelevant to moral responsibility. This paves the way for a new structuralist account that makes clear what it takes to be the diachronic ownership condition (which is normally taken to be personal identity) and the locus of moral responsibility (which is normally taken to be 'whole' person), and helps to alleviate the intuitive unease many have with respect to structuralism.

Keywords Manipulation · Moral responsibility · Historicism · Structuralism · Personal identity · Compatibilism

1 Introduction

Manipulation and moral responsibility have recently come to the forefront of the free will and moral responsibility debate. Incompatibilists have used manipulation *arguments* in attempts to undermine compatibilism.[1] Manipulation *cases*, on the

[1] For example, Derk Pereboom's 'four-case manipulation argument' (2001, pp. 110–118).

B. Matheson (✉)
School of Social Sciences, The University of Manchester, Arthur Lewis Building, Manchester M13 9PL, UK
e-mail: benjamin.matheson@manchester.ac.uk

other hand, have been used to motivate historical compatiblism, or historicism, in virtue of being counter-examples to non-historical or structuralist compatibilism.[2]

Structuralists, such as Frankfurt (1971), claim that moral responsibility depends only on the agent's psychological structure at the time of action. Given the nature of structuralism, it seems conceivable that an agent—such as a nefarious neuroscientist—could instil the appropriate psychological structure into another agent via, for example, brainwashing. Many find it intuitive that in such cases the manipulated agent is not morally responsible, given how she obtained her psychological structure; hence, structuralism ought to be rejected. Frankfurt (1975, 2002) stands firm and bites the bullet when faced with such cases—a stance which historicists find unpalatable. Thus, historicists motivate their position by claiming that they avoid counter-intuitive implications of structuralism.

I have two goals in this paper. My first is to show that historicism is not automatically preferable to structuralism. I argue that historicism has just as counter-intuitive implications as structuralism, thereby undermining the motivation for historicism. My second is to show that structuralism can be rendered much more plausible by appreciating the relationship between moral responsibility and personal identity. I sketch an alternative structuralist account, which I argue should help to alleviate the intuitive unease that many have with respect to structuralism. At the very least, it shows that the structuralist's bullet-biting stance is much more principled than many have thought.

I set the scene with preliminary issues in Sect. 2–4, and in Sect. 5–6 I argue that historicists must either commit to a psychological or a non-psychological approach to personal identity, but that committing either way comes with counter-intuitive implications. In Sect. 7 I end by sketching an alternative structuralist account which I argue renders it plausible that brainwashed agents are morally responsible.

2 Structuralism

Compatibilists must choose between two positions: historicism and structuralism. As I said earlier, structuralists hold that only factors that obtain at the time of action are relevant to an agent's moral responsibility. Frankfurt's (1971) structuralist account is a useful starting point. He claims that in order for an agent to be morally responsible, the hierarchy of her will must be in order. Roughly, this means that an agent's effective first-order desires (those desires which move her to action) must conform to her second-order volitions (her desires about which first-order desires she wants to move her action).

Frankfurt's structuralism is thought to succumb to counter-examples in the form of manipulation cases. These cases involve an agent who is covertly manipulated (often via brainwashing) into performing an action *A* whilst still satisfying the Frankfurtian conditions on moral responsibility. Given the manipulation, it seems

[2] In this paper, 'structuralism' will refer to all non-historical compatibilist positions. So, it will include both mesh theories, such as Frankfurt's (1971) hierarchical view, and a non-historical 'reasons-responsive' view.

intuitively plausible that the agent is not morally responsible for *A*, and so Frankfurt's conditions are inadequate. Frankfurt 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.
>
> (Frankfurt 2002, p. 28)

Frankfurt evidently thinks that certain sorts of manipulation are not a threat to moral responsibility: if manipulation provides an agent with a new character, that agent can subsequently be morally responsible for the actions that stem from that character. In Sect. 7 I sketch account that takes its cue from Frankfurt's bullet-biting stance.

Historicists, however, have been less than impressed with Frankfurt's response. Alfred 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, p. 294). And John Martin Fischer voices similar concerns when he says that 'The moral I draw from [manipulation 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, p. 209). Although historicists may not accept Frankfurt's bullet-biting response, I want to defend it and, in doing so, structuralism generally. The first part of my strategy for doing so will be to show that historicists have bullets to bite that are equally unpalatable.

3 Motivating historicism

In this section I consider one variant of the manipulation cases that Mele (1995, 2003, 2006, 2009a, b, 2013) uses to motivate historicism. These cases all have the same structure: an agent is provided with a new 'system of values' (which produces desires that lead to action), and then the agent acts as a result of this new value system. Since each of Mele's cases share the same structure, my arguments against one variant will apply to all his variants.

Consider the following case, which lacks manipulation:

> 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, p. 27). When he kills, he does so "because he wants to do it" (Frankfurt 2002, p. 27), and "he identifies himself with the springs of his action" (Frankfurt 1988, p. 54). When he was much younger, Chuck enjoyed torturing animals, but he was not 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 2013, p. 169)

This sort of case is then compared with a case in which an agent is manipulated 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 and 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, p. 27). Furthermore, the desire is "well integrated into [her] general psychic condition" (Frankfurt 2002, p. 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, p. 54), and she kills George "because

> [she] wants to do it" (Frankfurt 2002, p. 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.
>
> (2013, pp. 169–170)[3]

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.[4] 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 moral responsibility. This case appears to be a counter-example to not only Frankfurt's structuralism but all possible structuralist accounts, since whatever features of an agent's psychological structure that structuralists might point to as being relevant to the agent's moral responsibility can presumably be met in Brainwashed Beth. Hence this case appears to motivate historicism.[5]

4 Historicism and personal identity

It seems, then, that historicists have a counter-example to structuralism. However 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 John Locke:

[3] Note that in these sorts of cases the fact that Beth is manipulated by other agents isn't important. 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 (Mele 2009a, p. 168, n. 11).

[4] 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 this paper.

[5] Fischer and Ravizza (1998, pp. 195–196) argue that so-called 'tracing cases' also motivate historicism. These are cases in which an agent's moral responsibility for a particular action B *derives* from (or traces back to) an earlier action *A* for which the agent is *directly* responsible. The paradigm case in the literature involves an agent who drives whilst drunk and who is responsible whilst drunk because they were in control when they got drunk. However, I follow McKenna (2012b, pp. 266–267) in thinking that the debate between structuralists and historicists is over direct moral responsibility and that it is unproblematic for structuralists to accommodate the distinction between direct and derivative moral responsibility. There are also other manipulation cases which are problematic for structuralism, such as those featured in Pereboom's (2001) four-case argument, but each of these cases is also problematic for historicism; so I will not consider them here.

> 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 the night man would not be two as distinct persons as Socrates and Plato?
>
> (1690/1975, p. 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 non-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.[6]

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 for moral responsibility, 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.[7]

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 does this undermine the case for historicism? Well, I suggest that the main 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 manipulated

[6] 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.

[7] This case comes from the television programme *Twin Peaks*.

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 moral responsibility. Hence if pre- and post-Beth are numerically distinct, then Mele's case fails to demonstrate that structuralism fails; and therefore historicism is without motivation.

Unfortunately, however, in an earlier work, 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, p.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, p. 175, n.22)[8]

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' accounts on the one hand and psychological accounts on the other. Thus, we can elide the question whether Mele thinks the 'more in common' is

[8] In this and later quotes Mele is actually talking about different characters. The cases in which those agents 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, p. 176, b, p. 465).

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.[9] 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; and 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 will 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 manipulation cases are supposed to provide.

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? Pre-theoretically, 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

[9] 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.

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. But historicists cannot question this assumption, because 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 an incoherent 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 Sect. 4 above, 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 agent to have a history which is *hers*; and facts about *which* history is the agent's depend upon facts about personal identity. An agent *A's* history is the period in which there is an agent 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 agent does indeed require a history—specifically, they argue that an agent must have 'taken responsibility' for the mechanisms that lead to action. Taking responsibility requires that an agent see herself as the source of her actions and recognise that she is an apt target for the reactive attitudes (Fischer and Ravizza 1998, p. 230). But other historicists, such as Mele (1995, 2006) and Haji (1998), defend a *negative* historical condition on moral responsibility. This requires that an agent *lack* a history of particular sort—namely, one which includes responsibility-undermining factors. As such, Mele and Haji hold that moral responsibility does not require that

an agent have a 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, p. 155; my emphasis)

Practically unsheddable values are those values that an agent cannot modify or attenuate over a particular period; they are a deeply entrenched part of a person (Mele 1995, p. 153). According to Mele, when an agent 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 agent's own steam'—that is, without interference from external forces—that an agent can then be morally responsible for actions that stem from such values. This certainly sounds like a *positive* historical condition on moral responsibility: a condition that requires the agent 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 agent—at least one who acts from unsheddable values—requires a past.[10]

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 agent with 'unsheddable values'—that is, values which an agent 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

[10] For further support that historical conditions must have a positive strand, see McKenna (2012a, pp. 167–169).

could endorse is Olson's (1999) biological approach,[11] according to which persons are essentially human 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.

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 decide to 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 psychologically manipulated, and has a system of values which he accepts as his own. It seems that Chuckie is just as

[11] I will focus upon the biological approach, and in particular Olson's (1999) version of it, 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.

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 moral responsibility. 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, p. 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 *agent*'s history and what doesn't, which in turn establishes whether or not the agent 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 the previous section 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 no better a position than the structuralist; hence historicism is without motivation.

7 What about Beth?

If all this is correct, then there's no reason for compatibilists to 'go historical'—to use Mele's (2006, p. 176) words. The scales simply do not tip in favour of historicism. But what about Beth? Structuralists still have to incur the cost of biting the bullet in saying that she is morally responsible despite her brainwashing. But so what? Remember what Frankfurt says:

> 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, p. 28)

Since post-manipulation Beth certainly acts like a nasty piece of work in accordance with her will, why not treat her like a nasty piece of work? The structuralist can stand fast and accept that she is blameworthy for killing her neighbour, George. Of course many find this counter-intuitive, but it is no more counter-intuitive than thinking that either post-Chuck or Chuckie is *not* morally responsible. Of course, the structuralist would be in a much stronger position if there was some deeper rationale for the Frankfurtian bullet-biting stance that it seems she must take. Below I sketch an alternative structuralist account that aims to illuminate this deeper rationale.

One quick observation is in order before we proceed. There are (at least) two perspectives from which philosophers discuss moral responsibility. One is from the perspective of the free will debate. From this perspective philosophers often aim to provide the conditions which must obtain for an action to belong to an agent in the sense that they can, *at the time at which they perform the action,* be held responsible for it. We can call these the conditions of *synchronic ownership*. The disagreement between historicists and structuralists is over synchronic ownership: historicists claim there is a diachronic condition on synchronic ownership, whilst structuralists deny this. Nonetheless, both sides are interested in what makes the case that an agent is morally responsible for an action when she performs it. The other perspective comes from the personal identity debate. Philosophers working from this perspective often aim to provide the conditions which must obtain for an agent at t_2 to be morally responsible for an action performed at t_1. We can 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 person who committed the crime was *at the time* responsible (e.g. they were not sleepwalking or acting under extreme coercion), but also that the person in the dock bears the right relation to the person 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 agent at t_2 is the *same person* as the agent who

performed the free action. The debate between philosophers working from this perspective is thus normally a debate about the criteria for sameness of person.

The common assumption, then, is that diachronic ownership is a simply matter of personal identity. This assumption seems prevalent in the free will side of the moral responsibility literature because very little is said on the question of diachronic ownership; and it is this assumption, I shall argue, that is largely responsible for the allegation of implausibility against the structuralist position when it comes to manipulation cases. To see why structuralism isn't as implausible as many have supposed, I must first argue that numerical identity is not this diachronic ownership condition. I will show that changing this condition leads to a change in the *locus* of moral responsibility and this is central to making structuralism plausible. In this endeavour I follow in the footsteps of Parfit (1984), Schechtman (1996), Olson (1999), and Shoemaker (2012). It seems to me that the following sort of case is all that is required to make clear that personal identity (qua numerical identity) is not the diachronic ownership condition:

> *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 1,000 years. Some attribute this to the unique conditions on the 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 1,000 years old (call this temporal 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 connections with Clive-20. Indeed, Clive has gone through many different characters during his lengthy life.[12]

Suppose that Clive-20 performed action *A* and in doing so satisfied all the appropriate synchronic conditions on moral responsibility. 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 clear to me that he is 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 temporal slices shares more than 50 % of the psychological connections from the previous day) and biological continuity (that is, each of his temporal 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 personal identity were the diachronic ownership condition, simply seems

[12] This case is inspired by one of Parfit's (1984, pp. 302–303) cases. Parfit also uses 'fission' cases to support his claim that personal identity doesn't matter. I'm also uncertain whether personal identity doesn't matter to all practical concerns, but it seems clear to me that it doesn't matter to moral responsibility. And I think that the case can be made with more 'real life' cases of gradual character change, but it is much clearer with cases like Clive-1000.

preposterous.[13] Cases like Clive-1000 show that the diachronic ownership condition cannot be a transitive criterion. A criterion of personal identity must be transitive; thus, personal identity is not the diachronic ownership condition.

We must, then, find a replacement diachronic ownership condition. The implication of the Clive-1000 case is that any plausible diachronic ownership condition must not be a transitive relation. Although there are a few different possibilities for such a condition, in the following I will only consider one option.

I propose that the diachronic ownership condition is *narrative coherence*. This condition is found in many narrative views—in particular Schechtman's (1996) *narrative self-constitution view*. On the narrative view, an agent is morally responsible for a past action to the extent that the action coheres with the agent's self-told narrative.[14] An agent—call him 'Barry'—is morally responsible for any actions which cohere with his self-told narrative. Note that this narrative is more than likely implicit—Barry does not need to consciously organise his life in narrative form. It just has to be the case that his life takes this form to reasonable degree. The idea is simply that actions continue to belong to Barry to extent that they are *intelligibly* part of his self-conceived life story. Self-narratives must also cohere with *reality*; otherwise an agent might appropriate the action of another agent into her narrative. For example, Sally might be delusional and include an action of Barry's in her narrative. But Sally would not be an appropriate target of moral responsibility for this action because she is so out of touch with reality with respect to that action.

Notice that an outcome of accepting that numerical identity is not the diachronic ownership condition is that we must also change our view about what the locus of moral responsibility is. If we take numerical identity to be the diachronic ownership condition, then the locus of moral responsibility will be the 'whole' person—that is, the person across her entire existence—because ownership of actions remains with that person during her entire existence. If Barry performed a particular free action *A*, then he will always be morally responsible for *A*. But once we accept that narrative coherence is the diachronic ownership condition, we must also change our view about what the locus of moral responsibility is. It follows that the *narrative self* is the locus of moral responsibility.[15] So, Barry remains morally responsible for *A*-ing

[13] It has become common in recent literature to claim that an agent can *be* responsible but not be legitimately *held* responsible (e.g. Fischer 2006; Smith 2007; McKenna 2012a, b). An objector might claim that Clive-1000 is responsible for *A*-ing, but he cannot be *held* responsible—perhaps due to his character change. I lack the space to explain fully why I find this strategy problematic, though I will say this. When we say that an agent *is* responsible for an action it seems to me that this entails that this agent can *in principle* be held responsible—in other words, there is a possible agent who could legitimately praise or blame the agent in question. However, I find that there is no possible agent who could justifiably praise or blame Clive-1000 for *A*-ing. Hence, Clive-1000 *is not* morally responsible for *A*-ing.

[14] As this is only a sketch, there are certain worries with narrative views that I will not attempt to address, including the worries that narrativity is too subjective and that narrative views are unable to deal with our tendency to hold individuals morally responsible for 'out of character' actions.

[15] I leave undeveloped for the moment what narrative selves *are*. One possibility is that they are a kind of office that a person holds for a particular period of her life. This is in much the same way that one person can hold two different offices (such as being Chancellor and being Prime Minister) in her lifetime.

during the period in which he is the narrative self who *A*-ed. But if Barry ceases to be that narrative self, then he will no longer be morally responsible for *A*-ing. This is why Clive-1000 is not morally responsible for Clive-20's actions: Clive-1000, despite being numerical identical to Clive-20, is a different narrative self to Clive-20. Of course, most people will in fact continue to be the same narrative self throughout their life, so they will continue to be responsible for all their actions. We can then employ structuralist conditions (such as Frankfurt's or an account of reasons-responsiveness) to cover the synchronic ownership condition, which tells us which actions are free and which are unfree.[16]

With this account of structuralism in hand—one which makes clear what the diachronic ownership condition and the locus of moral responsibility are—we can see why, on this view, it is perfectly acceptable to think that post-manipulation Beth ($Beth_{V2}$) is morally responsible. Before starting, let's consider a feature of the Brainwashed Beth case I've so far overlooked. In order to support his claim that Beth is not morally responsible, Mele asks us to suppose that a day after George is killed that Beth regains pre-manipulation Beth's ($Beth_{V1}$'s) system of values—that is, the values of a kind and an innocent person. So for most of her life Beth—that is, the whole person—has been kind and innocent, apart from that day when she decided, in line with her values of the day, that she should kill George. On this way of viewing things, it's not surprising that many judge that Beth is not morally responsible, because Beth's dominant narrative self—which she is for almost all of her existence, aside from the 24-hour blip when she murders George—is clearly not morally responsible for killing him. However, the important question is whether $Beth_{V2}$—the narrative self who murders George—is morally responsible.

So, the question 'is *Beth* morally responsible?' is ambiguous; we need to know who 'Beth' refers to. Thus, we must first establish whether $Beth_{V1}$ (that is, Beth before and after the manipulation period) and $Beth_{V2}$ (that is, Beth during the manipulation period) are the same narrative self. It seems clear that they cannot be: $Beth_{V1}$ has the system of values of a kind and innocent person and $Beth_{V2}$ has the system of values of a serial killer; there is simply no coherence between $Beth_{V1}$'s narrative and $Beth_{V2}$'s narrative. Given that $Beth_{V1}$ and $Beth_{V2}$ are different narrative selves, this means that there is no transmission of moral responsibility between them. So if $Beth_{V2}$ is morally responsible for something, that doesn't mean that $Beth_{V1}$ also must be. This is similar to the Bob/Leland case I discussed in Sect. 4—just because Bob is morally responsible for a particular action, it does not follow

Footnote 15 continued
Usually a person will only be one narrative self during her whole life, but sometimes one person can be two narrative selves over the course of her life. Cf. Olson (1999, pp. 66).

[16] It seems to me that structuralist conditions need to be supplemented with a coherence condition of their own—for example, coherence between values. This coherence can then be used to explain the notion of 'identification'. Note that the view I'm sketching will have a coherence constraint on *both* the synchronic and diachronic ownership conditions. Thus, for an action to be a free action it must be that the agent is coherent at the time of action; and for an agent to be held responsible for past free action, that action must continue to cohere with the agent's narrative.

that Leland will also be morally responsible for that action. Of course, Bob and Leland are numerically distinct so it is uncontroversial that Bob's moral responsibility does not transmit to Leland. But it seems plausible that if the locus of moral responsibility is not the person, but rather the narrative self, that moral responsibility will not transmit between narrative selves. Hence, Beth_{V2}'s moral responsibility does not transmit to Beth_{V1}'s when she returns the day after George is murdered.

Is Beth_{V2} morally responsible? There no longer seems to be much intuitive pull towards thinking that she is not. The worry with holding that Beth_{V2} is morally responsible is, I have claimed, the worry that Beth_{V1} (when she returns) will be blameworthy for Beth_{V2}'s actions. However, the account I have set out does not have this consequence. The fact that Beth_{V2} is morally responsible for killing George does not entail that the sweet and innocent Beth_{V1} (when she returns) will get the blame. Therefore, it seems perfectly fine to accept that Beth_{V2} *is* morally responsible. If this counts as biting the bullet, then so be it.

8 Conclusion

Historical compatibilism is motivated by the claim that it avoids the counter-intuitive implications of structuralism. I have shown that this is not the case, because historicists must bite the bullet when it comes to thought experiments inspired by those in the personal identity literature. Historicists must appeal to an account of personal identity to make sense of what counts as an agent's past and so must commit to a particular account of personal identity. This highlights a further strength of structuralism: since it focuses on purely non-historical factors it can accommodate any account of personal identity. I then argued that structuralists can explain away the intuitive unease that many have in accepting that agents like post-manipulation Beth are morally responsible. This involved arguing that the locus of moral responsibility is not the person (in the strict sense), but rather the narrative self. Hence, a person is only morally responsible in virtue of being a particular narrative self. The key element of this view is a narrative coherence condition. With this condition in place, structuralists can make sense of the fact that post-manipulation Beth is morally responsible. Admittedly, this needs to be fleshed out in much greater detail but that is a task for another time. Although my defence ends in what may seem like bullet-biting to some, what I have shown is that structuralists can stand tall whilst enjoying the taste of lead. Indeed, given that historicism is without motivation, compatibilists have no reason not to be structuralists. I hope that the new structuralist account that I have sketched makes this a happy transition and that we can all join forces in thumping the table—Frankfurt-style—in future debates with incompatibilists.

Acknowledgments Author would like to thank the following people for comments on early and predecessor versions of this paper: Al Mele, Ann Whittle, Tim Bayne, Michael McKenna, Joel Smith, John Fischer, an anonymous reviewer for this journal, and the graduate community at the University of Manchester. Special thanks to Natalie Ashton. And extra special thanks to my supervisor, Helen Beebee, for countless comments on every version of this paper.

References

Fischer, J. M. (1994). *The metaphysics of free will*. Oxford: Blackwell.
Fischer, J. M. (2006). *My way*. Oxford: Oxford University Press.
Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control*. Cambridge, MA: Cambridge University Press.
Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy, 68*(1), 5–20.
Frankfurt, H. (1975). Three concepts of free action. *Proceedings of the Aristotelian Society, 45*, 113–125.
Frankfurt, H. (1988). *The importance of what we care about*. Cambridge, MA: Cambridge University Press.
Frankfurt, H. (2002). Reply to John Martin Fischer. In S. Buss & L. Overton (Eds.), *Contours of agency: Essays on themes from Harry Frankfurt*. London: MIT Press.
Haji, I. (1998). *Moral appraisability: Puzzles, proposals, and perplexities*. Oxford: Oxford University Press.
Locke, J. (1690/1975). Of identity and diversity. In J. Perry (Ed.), *Personal identity*. London: University of California Press.
McKenna, M. (2012a). Moral responsibility manipulation arguments, and history: Assessing the resilience of nonhistorical compatibilism. *Journal of Ethics, 16*(2), 145–174.
McKenna, M. (2012b). Defending nonhistorical compatibilism: A reply to Haji and Cuypers. *Philosophical Issues, 22*(1), 264–280.
Mele, A. (1995). *Autonomous agents*. Oxford: Oxford University Press.
Mele, A. (2003). Contours of agency: Essay on themes from Harry Frankfurt. *Australasian Journal of Philosophy, 81*(2), 292–295.
Mele, A. (2006). *Free will and luck*. Oxford: Oxford University Press.
Mele, A. (2009a). Moral responsibility and agents' histories. *Philosophical Studies, 142*, 161–181.
Mele, A. (2009b). Moral responsibility and history revisited. *Ethical Theory and Moral Practice, 12*, 463–475.
Mele, A. (2013). Manipulation moral responsibility, and bullet biting. *Journal of Ethics, 17*, 167–184.
Olson, E. (1999). *The human animal*. Oxford: Oxford University Press.
Parfit, D. (1984). *Reasons and persons*. Oxford: Oxford University Press.
Pereboom, D. (2001). *Living without free will*. Cambridge: Cambridge University Press.
Schechtman, M. (1996). *The constitution of Selves*. Ithaca: Cornell University Press.
Shoemaker, D. (2012). Responsibility without identity. *Harvard Review of Philosophy, XVIII*, 108–132.
Smith, A. M. (2007). On being responsible and holding responsible. *Journal of Ethics, 11*(4), 465–484.
Swinburne, R. (1984). Personal identity: The dualist theory. In R. Swinburne & S. Shoemaker (Eds.), *Personal identity*. London: Blackwell.
Thomson, J. (1997). People and their bodies. In J. Dancy (Ed.), *Reading parfit*. Oxford: Blackwell.