

Time and Moral Responsibility

A Thesis Presented to the
Chair of Examiners for the DPhil in Philosophy

Garrett Lam
Magdalen College

Abstract

Title: Time and Moral Responsibility

Candidate: Garrett Lam

College: Magdalen College

Degree: DPhil in Philosophy

The central theme of the thesis is how time—namely, what has happened in the past, what has happened in one’s history—relates to moral responsibility. The most common way this is discussed in the compatibilism literature is how someone’s history is relevant to whether he can be morally responsible. Non-historicists believe moral responsibility only depends on what state a person is in when he acts—what his causal springs are—and not how he came to be in that state. Historicists believe that moral responsibility also depends on how we come to possess our causal springs, and in the first part of the thesis I argue for a new type of historicist account of moral responsibility over existing versions.

What has happened in one’s history might not just affect whether he is responsible for what *is* done; it also might affect whether he is responsible for what *was* done. The second part investigates how historical considerations bear on whether a person at some time is responsible for what a person at some previous time did, a field which is little explored because most, but not all, authors assume the criterion is personal identity. I argue that alternatives to the personal identity criterion, namely those invoking psychological connectedness, face serious complications.

There is no literature to my knowledge on how positions we take on responsibility for what is done have implications for the plausible positions we could take on responsibility for what was done. In the third part of this thesis, I explore how these two ways in which history might affect moral responsibility relate to one another, arguing that of the permutations of historicist and non-historicist accounts with personal identity and psychological connectedness accounts, only historicist accounts of moral responsibility that require personal identity seem to lack serious complications.

Approximate Word Count: 70,853

Table of Contents

Title Page.....	i
Abstract	ii
Table of Contents	iii
 PART I: The Synchronic Conditions of Moral Responsibility	
CHAPTER I: Introduction	1
1.1 The Guises of Compatibility: A First Attempt to Avoid Verbal Disputes	1
1.2 Moral Responsibility: A Second Attempt to Avoid Verbal Disputes	3
1.3 Going Historical.....	4
 CHAPTER II: The Historicist Landscape	 11
2.1 Clarifications on Meaning and Method	11
2.2 Unsheddability.....	12
2.3 Pastless People.....	15
2.4 Positive and Negative Historicism	16
 CHAPTER III: Positive Historicism.....	 19
3.1 Fischer and Ravizza	19
3.2 McKenna	21
3.3 Childhood Values and First Free Acts.....	25
3.4 Unsheddability, Reconsidered.....	27
3.5 McKenna and Positive Historicism, Defended.....	30
3.6 Optimism Toward Positive Historicism, Reconsidered	35
 CHAPTER IV: Negative Historicism	 40
4.1 Mele.....	40
4.2 Innate and Early Childhood Values	42
4.3 Haji and Cuypers.....	44
4.4 The Problem of Arbitrary Differences	48
 CHAPTER V: Operant Historicism	 54
5.1 Operativity.....	54
5.2 What Mele Did Consider: Can Manipulated Agents Regain Their Freedom?	56
5.3 The Road to Responsibility: Why This Isn't Just Positive or Negative Historicism	59
5.4 Comparing Operant Historicisms	65
5.5 What Mele Didn't Consider: How Radical Reversal Affects Radical Restoration.....	69
5.6 Fine Tuning Operativity.....	76
 PART II: The Diachronic Conditions of Moral Responsibility	
CHAPTER VI: The Other Side of History	84
6.1 Blame over time	84
6.2 Blameworthiness and Blame	85
6.3 The Link to Personal Identity	87
 CHAPTER VII: The Sufficiency of Personal Identity.....	 91
7.1 Theories of Personal Identity.....	91
7.2 Incapacitated Agents	92
7.3 Can We Modulated Blameworthiness with Different Actions?.....	95
7.4 Can We Modulated Blameworthiness with Different Psychology?	97

CHAPTER VIII: The Necessity of Personal Identity	104
8.1 Pressure on the Necessity Claim	104
8.2 The Psychological Connectedness Account	107
8.3 Freak Coincidences and Post-Hoc Retributions	111
CHAPTER IX: Causation and the Psychological Account	116
9.1 Responsibility with Any Cause of Continuity	116
9.2 Freak Coincidences and Post-Hoc Retributions, with Causation	121
9.3 Normal/Narrow Causes	127
9.4 The Reasons-Revision	136
9.5 A Universally Troublesome Case for Psychological Connectedness Views	140
PART III: Intersecting the Synchronic and Diachronic Conditions	
CHAPTER X: The General Psychological Account and Synchronic Conditions	143
10.1 Permutations	143
10.2 The Psychological Account and Manipulated Agents	144
10.3 The Psychological Account and Instant Agents	147
10.4 Non-historicism to the Rescue?	150
10.5 Moral Scapegoating	151
CHAPTER XI: The Reasons-Revision and Synchronic Conditions	156
11.1 Causation to the Rescue?	156
11.2 The Reasons-Revision and Historicism	161
11.3 The Re-emergence of Re-emergent Autonomy	165
CHAPTER XII: Personal Identity and Synchronic Conditions	172
12.1 Can Identity Fare Better?	172
12.2 Personal Identity and Non-historicism	172
12.3 Personal Identity and Historicism	176
References	181

PART I: The Synchronic Conditions of Moral Responsibility

CHAPTER I: Introduction

1.1 The Guises of Compatibility: A First Attempt to Avoid Verbal Disputes

If you tell me you're a determinist, I know exactly what you believe: at any instant, there is exactly one physically possible future.¹ I'll think of atoms bouncing around in a mechanistic fashion. The future unfolding in a clockwork manner. Each event the physical realization of something that was causally set in stone billions of years before. But if you tell me you're a compatibilist, you'll get a follow up question: compatibilist *about what*? Presumably, you believe something is compatible with the truth of determinism. But is it moral responsibility? Free will? Punishment? The term seems to get thrown around to cover all of these.² What's more, it is often used to link some of them. For instance, Michael McKenna defines "free will as the unique ability of persons to control their conduct in the strongest manner necessary for moral responsibility."³ Derk Pereboom, following Alfred Mele,⁴ takes "'free will' to refer to the strongest sort of control in action required for a core sense of moral responsibility."⁵

These authors all recognize that being responsible (uncontroversially) requires some sort of control by the person who is said to meet the condition, and consequently use 'free will' to refer to whatever sort of control is *necessary* for moral responsibility. This is a linguistically expedient practice in so far as it lets us say things like "Beth does not freely A" when we want to say that Beth does not satisfy the conditions necessary for being morally responsible when she A's. (We can thus avoid the awkward "Beth does not responsibly A" which has too many misleading connotations from popular usage,⁶ or the utterly verbose "Beth does not A such that she is morally responsible for it.") But it is a philosophically risky practice in so far as it assumes that if we lack free will we also lack moral

¹ Following van Inwagen (1983), 3.

² For discussions of compatibilism between determinism and punishment, see Greene and Cohen (2004) and Morse (2004).

³ McKenna (2012b), 147.

⁴ Mele (2006).

⁵ Pereboom (2014), 2.

⁶ Or "Beth does not morally responsibly A," which offsets its increase in clarity with a boost in awkwardness.

responsibility. Now, this is perfectly kosher if we simply define free will as something necessary for moral responsibility. But many philosophers believe that free will is important for non-normative reasons, and stipulating free will in such a way seems uncharitable toward the so called semicompatibilist positions held by people like Fischer, who think moral responsibility is compatible with determinism but are agnostic about whether free will is, and thus hold that one can consistently believe moral responsibility is compatible with determinism while free will is not.⁷ Our definitions shouldn't automatically preclude a potentially viable position. Of course, what could be going on here is that Fischer believes free will might be incompatible with determinism in a way that is consistent with McKenna believing free will is compatible with determinism (since both believe in moral responsibility)—but then we run the risk of smearing a substantively important debate into a merely verbal dispute.⁸

I will not be arguing about whether we should understand free will as that which is necessary for moral responsibility or as something else (like the ability to do otherwise or to act from alternative possibilities, the lack of which, some argue, still allows for moral responsibility). I do not know how to do so, and sometimes I'm not even sure what would come from answering such a question. Rather, I will be focusing within the realm of moral responsibility's compatibility with determinism, leaving open whether we should understand free will in a specific way. The fact that there is, as I have mentioned, no non-misleading or non-awkward way of adverbializing moral responsibility means that I will follow convention and use formulations like "Beth does not freely A/Ann is free/X is freedom-undermining" as shorthand for "Beth is not morally responsible for A-ing/Ann is morally responsible/X undermines moral responsibility." But I will only be making claims about responsibility, not free will.

⁷ Fischer (2014), 206-207.

⁸ See Chalmers (2011) for an overall examination of verbal disputes in philosophy. On pp. 530, Chalmers discusses how free will might be construed as a verbal dispute, though he believes the same might also apply to moral responsibility.

1.2 Moral Responsibility: A Second Attempt to Avoid Verbal Disputes

The risks of verbal disputes have still not been fully addressed. Since I will be discussing whether moral responsibility is compatible with determinism, we need a working definition of moral responsibility. The one I have in mind is the most common in the literature around free will and determinism where moral responsibility is a sort of status intimately connected to blameworthiness and praiseworthiness and the judgments of blame and praise. This is what Pereboom calls the “basic desert”⁹ sense: for an agent to be morally responsible for an action is for it to be his in such a way that he would deserve to be blamed if he understood that it was morally wrong, and he would deserve to be praised if he understood that it was morally exemplary. The desert is basic since the person deserves praise or blame simply because he knowingly did the act, in contradistinction to whether a person might deserve blame because it would educate him of its wrongness and prevent him from doing the act in the future, or would prevent others from doing it in the future. For instance, Daniel Dennett identifies as a compatibilist, and puts forward a conception of blame and punishment that is largely utilitarian. As Morris writes:¹⁰

[Dennett’s account] comes at a great price...he has defined the concepts of “free will” and “moral responsibility” in such a way as to eliminate any substantive difference between the “compatibilist” position he defends and the hard determinist position that philosophers typically understand as being substantively different from compatibilism...Thus, in redefining “free will” and “moral responsibility” in such a way that they could each apply to an agent whose actions are caused by purely unconscious processes, any disputes that Dennett would appear to have with hard determinists have become wholly verbal.

As I have stressed, I have no intention of determining whether a rose by any other name would smell as sweet. Compatibilism about moral responsibility, as I will use it in this thesis, will refer to the substantive claim that the truth of determinism is compatible with people deserving, in the basic desert sense, praise and blame.¹¹ And the purpose of this thesis is to get closer to determining whether or

⁹ Pereboom (2001), xx.

¹⁰ Morris (2009), 69.

¹¹ In structuring the thesis around this conception of moral responsibility, I am following the conventions around the dialectic (i.e. the historicist vs. non-historicist debate within compatibilism). It is worth pointing out that there is an

not this substantive claim is true.

1.3 Going Historical and a Brief Outline of Time

So far, so semantic. Now that we've bracketed verbal disputes to isolate what it means to be a compatibilist, we can articulate the compatibilist challenge. Compatibilism says that determinism is not a threat to moral responsibility. But of course we are not always responsible for our actions, and we readily take ignorance, coercion, and mental illness to subvert autonomy. So the compatibilist challenge is to articulate why even if all our actions have causal springs, only some causal springs are responsibility-preserving—why all actions are caused but only some causes undermine freedom.

What unites all compatibilist accounts are desiderata for the necessary psychological conditions one must satisfy to be morally responsible. These almost always include epistemic and control-based conditions. Regarding the former, we readily excuse the mother for accidentally feeding her child poisonous peanut butter if it looked completely normal but was toxic from a mistake in the manufacturing process (and so her ignorance was non-culpable). Control requirements vary, but might require that someone have the ability to understand and follow moral reasons.¹² Additionally, sometimes identification conditions are appealed to, perhaps requiring that someone identify with the desire that is driving his action.¹³ What is important to recognize about these desiderata is that all these responsibility-grounding elements could be read off of a time-slice snapshot of someone; they refer only to what the person is presently like and what the causal springs are, not how that person came to be that way or have those causal springs in the first place. A complete internal accounting of a person and his causal springs would tell us whether any of these requirements have been violated.

enormous amount of literature on different traditions and conceptions of moral responsibility, and many authors (see Shoemaker 2014) proffer pluralistic understandings of moral responsibility. Adjudicating between them (e.g. why a particular conception of moral responsibility within the Strawsonian tradition is the right one) exceeds the scope of this thesis, and hence I will follow the standards used by the writers I will be dominantly discussing.

¹² Watson (1999).

¹³ Frankfurt (1971).

If one of these requirements is not met when someone acts, he is not responsible for that action. But if none of them have been violated, can we definitively say that he is responsible? Here we come to an “in house” debate among compatibilists—a disagreement about whether moral responsibility is a *historical* phenomenon. Some properties, like color, are non-historical. The history of how a painting came to be has no bearing on whether it has some portion of a specific shade of blue. Compatibilists who are non-historicists believe that moral responsibility is like this; if two people act and are identical with respect to their non-historical properties, they must both be responsible or both act unfreely. But other properties are historical. Two paintings can be identical with respect to their structural properties, but only the one is a genuine Rembrandt if it was painted by the man himself while the other was machine-generated. The property of being-a-genuine-Rembrandt is clearly historical. Compatibilists who are historicists believe that even if two people act from causal springs that are identical with respect to their non-historical properties, we don’t know enough to determine whether either of them is responsible. We don’t just need to know what the causal springs of action *are*, we also need to know *how those causal springs came to be*.

Harry Frankfurt offers an illuminating assertion of the non-historicist position. After putting forward his own requirements that, essentially, require that someone identify with the desires that move him to perform some action, he writes:¹⁴

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

Frankfurt, especially at the time of writing, was by no means alone in his defense of non-historicism; compatibilists have historically been non-historicists.¹⁵ Yet, there seems to be shift toward historicism

¹⁴ Frankfurt (1998), 54.

¹⁵ Hobbes and Hume articulate the traditional compatibilist conditions (see Russell (1998), for an overview). Also see McKenna (2012a), 264.

in the face of an enduring gadfly: manipulation arguments. Manipulation arguments against non-historicism all share similar structures: a case is presented in which an agent is subject to intense psychological manipulation—involving the covert implantation of seriously opposing values, often by evil neuroscientists—but is manipulated in a way that preserves non-historicist conditions. It is expected that the thought experiment elicits a ‘non-responsible’ intuition, revealing a deficiency in non-historical accounts. Now, these arguments can take particular or general forms. Particular manipulation arguments are meant to show deficiencies of specific non-historicist accounts: an agent is manipulated in a way that satisfies all the compatibilist conditions of the specific account (e.g. the agent is manipulated to identify, in Frankfurt’s sense, with his action). In so far as one judges the agent to nevertheless lack moral responsibility, the specific account is deficient. General manipulation arguments draw a picture of an agent that any plausible compatibilist account will deem free. Next, some other agent is manipulated to become a non-historical duplicate of that free agent. In so far as one judges the manipulated agent to nevertheless lack moral responsibility, *any* non-historical account will be deficient.¹⁶

¹⁶ It is important to note that manipulation arguments are commonly given as arguments for *incompatibilism*, rather than historical compatibilism, so it is worth situating manipulation arguments for historical compatibilism within the broader class of manipulation arguments. Following King (2013), I understand the structure of manipulation arguments for incompatibilism as:

- (M1) The manipulated agent is not morally responsible for acting on his implanted psychological states, despite satisfying compatibilist conditions on moral responsibility.
- (M2) There is no difference relevant to moral responsibility between the manipulated agent’s action and the actions of ordinary agents in a deterministic universe.
- (MC) Thus, compatibilist conditions on moral responsibility are insufficient. Compatibilism is false.

Since (M1) says that the manipulated agent is not morally responsible, and since (M2) says we should deliver the same verdict to the manipulated agent and an ordinary deterministic agent, there are two general strategies the compatibilist can take in resisting the manipulation conclusion (MC): what McKenna (2008) calls the ‘hard-line’ and ‘soft-line’ reply. If he takes the ‘hard-line’ reply and denies (M1), he admits that there are no responsibility-relevant differences but bites the bullet and maintains that, contrary to appearances, the agent *is* morally responsible. If he takes the ‘soft-line’ response and denies (M2), he agrees that the agent is not responsible but argues that there is something unique to manipulation that undermines responsibility. With this in mind, it should be clear how historicist arguments square with incompatibilist arguments. Both historicists and incompatibilists will sustain (M1), believing that non-historicists will implausibly bite the bullet. But the historicist compatibilist and the incompatibilist will disagree about (M2); the historicist will argue that there are *historical conditions* on moral responsibility that the manipulated agent doesn’t satisfy, while the incompatibilist will argue that there is no such difference between the manipulation and ordinary determinism.

What does a general manipulation argument for historicism look like? Consider Mele's case of Ann and Beth. Both are philosophy professors, but Ann is far more committed to philosophy. She spends the majority of each day researching, while Beth spends far less. The university dean wants Beth to be more productive, so recruits advanced neuroscientists to covertly implant Ann's values into Beth. The manipulation turns Beth into Ann's "psychological twin,"¹⁷ and her newly-engineered hierarchy of values causes her to drastically change her life to revolve around philosophy. Is Beth still responsible for her actions? A non-historicist believing that Ann is morally responsible will also say that Beth is since the two have become value twins, but Mele writes:

[Beth's] own considered values were erased and replaced in the brainwashing process...it is difficult not to see her now, in light of all this, as heteronomous—and unfree—to a significant extent in an important sphere of her life. If that perception is correct, then given the psychological similarities between the two agents, the difference in their current status regarding freedom would seem to lie in how they came to have certain of their psychological features, hence in something external to their present psychological constitutions. That is, the crucial difference is historical; free agency is in some way history-bound.¹⁸

There's the rub with non-historicism: if the dean's neuroscientists preserve non-historicist conditions, they can drastically alter Beth's entire character, completely bypassing her control capacities, eradicating and replacing her values without her consent or knowledge, and she would still be free. More generally, take any morally responsible murderer, thief, or arsonist; if non-historicism is true, clever neuroscientists could turn you into his psychological duplicate and make you responsible for murders, thefts, and fires. Historicists can try to argue a principled difference. Non-historicists have no choice but to bite the bullet.

Frankfurt, unequivocally non-historicist, is under no illusion of this, writing more recently, and most provocatively:

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

¹⁷ Mele (1995), 145-146.

¹⁸ Ibid.

change us radically, without thereby bringing it about that we are not morally responsible 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.¹⁹

For many, this is too cold and unforgiving an understanding of human agency. As Fischer writes, “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.”²⁰ And Mele echoes: “if compatibilists were to have nothing more attractive to offer than Frankfurt’s [non-historicist] view of moral responsibility and freedom, compatibilism would be in dire straits.”²¹ Perhaps this dismissal of non-historicism is premature, but it certainly justifies an investigation of the tenability of historical compatibilisms.

In the first part of this thesis, I do just that. In Chapter II I make several methodological comments and give an overview of the two broad classes of historicist accounts: positive accounts which require that an agent has a past of a certain sort; and negative accounts which require that an agent lacks a past of a certain sort. In Chapter III I focus on positive historicism. Positive historicism, I argue, faces a complication: it is no more plausible than a comparable non-historical account of moral responsibility because it makes it too difficult to take ownership for one’s causal springs. In Chapter IV, I then turn to negative historicism, which faces an equally severe complication: it has implausibly revisionist implications for who is and is not responsible because it makes it too easy to take ownership for one’s causal springs and consequently delivers responsible and non-responsible evaluations of agents who seem only arbitrarily different, and whom any sound account of moral responsibility would deliver symmetric judgments too. These two arguments suggest that positive and negative historicism fail by falling on opposite sides of how demanding a conception of authenticity

¹⁹ Frankfurt (2002), 28.

²⁰ Fischer (1994), 209.

²¹ Mele (2006), 172.

they require. Positive historicism seems to set the bar too high for what it means to authentically own one's values, resulting in an overly restrictive and so overly exclusive conception of autonomy; negative historicism seems to set the bar too low, resulting in an overly liberal and so overly inclusive conception of autonomy. However, there does seem to be a middle ground, and in Chapter V I propose a novel feature of historicism, what I dub operativity or operant historicism, that navigates between the complications its positive and negative cousins face and can be used to supplement either positive or negative historicism to increase their plausibility.

Part I only considers the conditions or blameworthiness *for performing an action*, and whether what has occurred in one's history is part of such conditions. However, there is another way that history can relate to moral responsibility, and that is the question of how *blameworthiness extends over time*, or how responsibility *transfers* into the future—how what an agent did in the past relates to his, or anyone's, responsibility in the present. This area of inquiry includes not only all of the consideration that bear on whether a person who freely does some action is responsible for that action at some point in the future, but more broadly how someone, even someone who is numerically non-identical to the prior responsible party, might retain or inherit responsibility for something that was done freely in the past, and that is the focus of Part II of the thesis. It is extremely natural to think that if I am responsible for doing something, I am responsible for having done it the next minute, the next hour, and the next day. And the same would seem true for subsequent days, and weeks. It does not seem like blameworthiness dissipates over time. Interestingly, there seems to be an implicit presumption about personal identity in the phrasing of this point. It seems obvious that personal identity is a necessary component of the conditions for preserving blame for some prior blameworthy act: how can *I* still be responsible for an act that wasn't even committed by *me*? And the notion that blameworthiness lasts forever also makes personal identity seem sufficient for responsibility over time.

In Chapter VI and VII, I consider the views that personal identity is not sufficient for moral responsibility. Some authors also believe it is not necessary, and argue versions of responsibility for past actions that depend on *psychological connectedness* between the evaluated agent and the past free actor. Such psychological connections themselves may have no causal requirements or they might have causal requirements (i.e. requirements on appropriate causal histories for how those psychological connections obtain) and in Chapters VIII and IX, I argue against the plausibility of both types of psychological connectedness accounts of moral responsibility for past actions. My arguments, if sound, do not deny that psychological connectedness plays a part in the correct account of moral responsibility, only that it supplements that view that personal identity is necessary for responsibility, rather than replaces it.

There is extensive literature around what I covered in Part I, which explores historicism. And there is a little literature around what I covered in Part II, which explores psychological connectedness vs. personal identity accounts of responsibility for past actions. But there is to my knowledge *no* literature around how, say, a commitment to non-historicism makes harder for one to plausibly assert a personal identity account, or other complications when permutating positions across the categories. In this final part of the thesis, I do this mix and matching. In Chapter X I argue that a non-causal version of the psychological connectedness account suffers from intersectional complications when paired with both non-historicism and historicism, as does, as I argue in Chapter XI, a causal version of the psychological connectedness account. In Chapter XII, I argue that the personal identity view also has interactive conceptual costs when conjoined with non-historicism, and only a pairing of the personal identity view with historicism seems to lack any obvious interactive complications.

CHAPTER II: The Historicist Landscape

2.1 Clarifications on Meaning and Method

Before continuing, it is important to make two clarifications. The first concerns one final verbal issue that might arise when distinguishing historicist from non-historicist accounts. Consider a man who freely gets drunk at a bar and drives home. He lacks even basic coordination, momentarily blacking out and hitting a pedestrian on the side-walk before waking up to an airbag in his face. Nearly everyone who believes that people are sometimes morally responsible will presumably want to uphold the man's culpability in hitting the pedestrian, but since he was essentially blacked out at the moment he hit the unfortunate chap, *at that time* he plausibly did not satisfy control conditions for moral responsibility. The textbook reply here is to agree, but maintain that his responsibility *traces* back to an act that he did freely do: drink in the first place, knowing he could have his decision-making impaired and drunk drive, which could have such consequences as hitting a pedestrian. He is then *indirectly responsible* for hitting the pedestrian, the responsibility for which *derives* from his being directly responsible for drinking. Now, since nearly all compatibilist accounts will invoke this sense of responsibility being historical, it is not helpful in distinguishing a substantive difference between historicism and non-historicism. Thus, my discussions on historicism will bear on actions for which one is *directly* morally responsible.²²

The second point relates not to the meaning of various terms, but the manner in which arguments are presented. Some strange characters inhabit the following pages. The ones I have in mind are agents, with the same degree of cognitive complexity as you and I and a rich set of (false) memories, who were created out of thin air and have no pasts. And god-like creatures who can create such past-less people. And nefarious neuroscientists who have the technology and malice to zap people's brains, thereby changing entire swaths of values to their polar opposites. Then we will move

²² See McKenna (2012b) for a similar discussion.

to people whose brains are divided and placed into new bodies, and those that live for hundreds of years. I will freely invoke such eccentric characters and outlandish cases with the intention of eliciting various intuitions to support and challenge various accounts of moral responsibility. But many philosophers distrust the contents of our intuitions as guides to the truth²³, especially those about far-fetched scenarios.²⁴ This is, moreover, particularly salient when it comes to blame, as one need not be a radical Strawsonian to think the concepts of responsibility have their home in particular kinds of distinctively human interpersonal dealings and relationships—in which case, how much could we learn from examples so far removed from normal life?²⁵ My comfort with using such thought experiments stems from the fact that I will be criticizing accounts of moral responsibility that were themselves asserted with the conceptual support derived from the very same types of thought experiments. In so far as someone invokes the concept of an instant agent to defend a historicist account of moral responsibility, so too can I invoke instant agents in challenging that account. If eccentric creatures can be used to support a theory, creatures that are no more eccentric can be used to attack it. Any claims of immunity from the very conceptual resources used to advocate for a position seem wide open to *tu quoque* accusations.

2.2 Unsheddability

While causal springs give an intuitive metaphor for the mechanisms by which an agent acts, it is far more common when discussing moral responsibility to talk of agents acting from or according to *values*, since our values determine or inform our conduct. Since such value-talk will figure prominently in much of what I discuss going forward, it is crucial to specify exactly what I mean. I adopt Mele’s terminology: an agent S “*thinly* values X at a time if and only if at that time S both has a positive

²³ Cappelen (2012).

²⁴ Wilkes (1998), Chapter 1.

²⁵ Thanks to Christopher Bennett and Cressida Gaukroger for pushing this.

motivational attitude toward X and believes X to be good.”²⁶ But there is a type of value that is particularly important when discussing cases of manipulation, what Mele calls an *unsheddable value*,²⁷ where a value is unsheddable for an agent (over some time t) if and only if, given the agent’s “psychological constitution,” eradicating or significantly attenuating the value over that time period “is not a psychologically genuine option.”²⁸ An agent with an unsheddable value over some time period is in some sense stuck with it, since he is unable to significantly mitigate the extent to which it plays a role in his actions.

Why does unsheddability bear so much relevance to historicist views? It is usually understood that when we say that a person is morally responsible for some action, we mean he *autonomously* possesses the values that figure into producing that action, or they are *authentic*. Conversely, manipulation undermines moral responsibility because it implants inauthentic values in a person so that he does not autonomously possess them when he makes choices on their basis. Thus, a fruitful avenue for exploring the conditions of moral responsibility is to ask when a person autonomously or heteronomously possesses his values. And since we cannot non-arbitrarily determine or alter our values without some pre-existing evaluative framework—moreover, since our character is shaped and influenced long before we are in any position to shape and influence that very character—it is common to understand autonomy in terms of *abilities* to control the values, desires, etc., that we already have.²⁹ For instance, Dworkin approximates autonomy as “a second-order capacity to reflect critically upon one’s first-order preferences and desires, and the ability either to identify with these or to change them in light of higher-order preferences and values.”³⁰ While what exactly constitutes the nature of ability

²⁶ Mele (1995), 116.

²⁷ Following convention, I leave out the ‘practical’ for brevity.

²⁸ Mele (1995), 153-154.

²⁹ Indeed, the basic Strawsonian argument against the possibility of moral responsibility, compatibilist and incompatibilist alike, is that we can’t act authentically, since we can’t choose without a choice-making mechanism and we cannot choose what choice-making mechanism we initially have—it’s given to us. See Strawson (1994).

³⁰ Dworkin (1998), 108.

is the subject of much debate, and one that we will not (and need not) explore in depth here, the basic idea is that some correct articulation of this ability will describe the capacity for critical reflection (of one's desires and values in light of higher order-ones) which seems good grounds for autonomous possession of those values. Mele writes:³¹

Suppose that *S*, who is possessed of a wealth of first- and higher-order pro-attitudes, has and exercises a potent capacity for critical reflection. Further, with respect to each of his pro-attitudes—of *any* order—he has [the proper] ability. On a robust construal of 'ability,' according to which ability has considerably more bite than mere logical possibility, *S*'s history (whatever it may involve) seemingly constitutes no bar to his possessing psychological autonomy (at least, if compatibilism is true)—provided that his abilities are suitably employed. *S* apparently would be the master of his values, principles, preferences, and desires: whether he identifies with them or changes them would be up to him. If [this] ability with respect to one's pro-attitudes (paired with suitable employment of the ability) *were not only sufficient for psychological autonomy but necessary as well*³², the [non-historicist] would carry the day.

Mele cashes out this ability in terms of sheddability, which is sufficient for psychological autonomy.

A sheddable value is, by stipulation, one for which an agent has the requisite amount of cognitive control over such that he is responsible for acting from it (with various other compatibilist criteria being satisfied); according to this understanding, no compatibilists should argue that the implantation of sheddable values—values whose strength an agent can appreciably diminish—is responsibility-undermining, since it seems implausible to think that if we are sometimes morally responsible, that responsibility could get subverted if a manipulator implants a desire to steal that is weak enough that I can actually eradicate it but, over the course of practical deliberation, nevertheless sustain and consequently act on.

On the other hand, unsheddable values that are injected into one's psychology seem like prime candidates for freedom-undermining interventions; as Mele notes, an unsheddable value is similar to an irresistible desire since the value is deep enough in an agent's evaluative perspective that he cannot attenuate the value's influence in the relevant practical deliberation. And, while the intuition will be

³¹ Mele (1995), 151.

³² Emphasis mine.

cashed out in far more detail in the proceeding sections, it is *prima facie* plausible that a person like Beth who is manipulated to possess an unsheddable value without her knowledge or consent is not responsible when it drives her actions in a way that is orthogonal to how she would have acted had she not been manipulated.³³ Thus, even if Beth's values are identical to Ann's with respect to each of their *abilities* to alter their character, the *way in which* Beth developed such equally entrenched values seems relevant to responsibility. The historicist's task is to spell out under what histories we can and cannot autonomously possess *unsheddable* values.

2.3 Pastless People

It is time to introduce another class of outlandish characters. Consider instant agents: beings with cognitive abilities as complex as yours and mine, but who do not develop such complexity over time—they are created that way. The first one we meet is McKenna's Suzie Instant³⁴, who was created instantaneously by a god and is psychologically indistinguishable from a normal 30-year-old. She has an enormous set of false memories about her (non-existent) past and false beliefs about how she came to have the values she does. Moreover, she satisfies [pick-your-favorite] non-historical account of moral responsibility, and is given a choice to A or B, where A would involve acting from an unsheddable value while B would violate it. When Suzie Instant A-s, is she responsible for it?

To help bolster our intuitions about such a strange scenario, McKenna recruits the same god who created Suzie Instant to go to a different (deterministic) world and create Suzie Normal, who develops in a normal zygotic manner in such a way that 30 years later she is a psychological twin (and non-historical duplicate) of Suzie Instant, and thus A-s. If anyone is ever responsible, Suzie Normal is. After all, she could share the exact same history as Ann in Mele's Ann/Beth case—who develops from a zygote during normal biological reproduction with no help from a powerful god—save the

³³ Of course, a value being unsheddable isn't sufficient for it being inauthentic; all authors discussed here implicitly or explicitly endorse a compatibilist account according to which we sometimes freely act from unsheddable values.

³⁴ McKenna (2004), 180.

intentions of an external creator, a factor that all sides agree is irrelevant to moral responsibility. But, as McKenna notes, it would be weird if Suzie Instant were unfree simply because her causal history “came compressed in a momentary package where Suzie Normal’s history chugged along over the course of thirty years.”³⁵

McKenna is hardly alone in his judgment of Suzie Instant’s autonomy. Mele’s even more bizarre minutelings are created by taking cloned human bodies and inserting some carefully constructed psychological profiles, yielding conscious beings who apparently can only live for a minute.³⁶ Mele has us imagine someone creating minutelings to kill a prisoner on a firing line. Mele reports: “I have not seen a convincing argument that would yield the judgments that my minuteling does not freely decide to shoot the prisoner and is not morally responsible for deciding to do that.”³⁷ Thus, let us for now say that it is *prima facie* plausible that instant agents are morally responsible.

2.4 Positive and Negative Historicism

We are now in a position to roughly sketch out the putatively two types of historicist accounts. While all historicists maintain *that* history matters for moral agency, they disagree about *why* history matters. One way of distinguishing the requirements of an agent’s history is whether such constraints are *positive* or *negative*:

Positive Historicism: An agent is morally responsible if and only if he has a history in which some event(s) has/have actually occurred.

Negative Historicism: An agent is morally responsible if and only if he lacks a history in which some event(s) has/have actually occurred.

The basic idea behind the distinction is that when we evaluate whether some sequence of events satisfies a given condition, depending on the condition we might need to use one of two strategies. If the condition specifies a positive constraint, then we need to scan his history and see whether anything

³⁵ Ibid, 181.

³⁶ Mele (2013), 162.

³⁷ Ibid. 165.

in it matches the criteria; if the condition specifies a negative constraint, we need to know that some criteria are not in it. An academic award that requires that one has been the examiner of a DPhil thesis places a positive historical constraint, since a candidate's eligibility would be affirmed by tracing the actual sequence of his academic tenure and verifying that the event of examining-a-DPhil has occurred. On the other hand, an academic position requiring that one is free from suspensions from any education institute places a negative historical constraint, since a candidate's eligibility would be affirmed by tracing the actual sequence of his academic tenure and verifying that the event of getting-suspended has never occurred.

It is important to cash this out in terms of events or series of events, rather than properties, since events over some time interval actually happen or they do not, and thus one cannot arbitrarily convert a negative constraint into a positive one in the same way one could turn the presence of a property into a negative condition by defining its absence in positive terms, or the absence of a property into a positive condition by defining its presence in negative terms. Hence, even if was-free-from-manipulation might have the ring of positive constraint since it spells out a property (of not being manipulated) that one must have rather than a property one must lack, this is really a negative constraint since, in event terms, it is equivalent to lacking a history in which there was manipulation.

We can appreciate why each of these two seemingly collectively exhaustive types of historicism has some appeal. There are at least two desirable characteristics of any account of moral responsibility. First, the account should meet certain principles. For instance, a common compatibilist intuition is that a responsible person must *take ownership* over his actions. Given that one is attracted to historical accounts of moral responsibility, he will likely find structural accounts of taking ownership, such as Frankfurt's, too thin. If this is true, he will probably require that someone has *done something* to take ownership prior to acting from those causal springs, and this is clearly a positive historical constraint that only a positive historicism can vindicate.

However, in addition to meeting certain principles, the account should ideally deliver the intuitive verdicts for various thought experiments, lest it transform into something too revisionary. For instance, as we have seen, many compatibilists find it intuitively plausible that instant agents like Instant Suzie are morally responsible for their actions immediately after creation. Though perhaps a compelling story could be told for debunking such intuitions or otherwise explaining why we should not trust them, this adds at least *pro tanto* support to whatever theory of moral responsibility holds such instant agents responsible for their initial actions. But instant agents, by stipulation, have no past, and so any historicist account that affirms their freedom must be a negative one.

It seems that the two types of historicism each capture one of two important compatibilist criteria, but not both. Are we at a dialectic stalemate with no additional conceptual resources other than an intuition tug-of-war? In the following two chapters I evaluate, respectively, the relative and absolute merits of positive and negative historicism. Both face severe complications. However, spelling out such complications puts us in a position to appreciate that the positive/negative distinction may not be as collectively exhaustive of historicism's resources as it has historically been taken to be.

CHAPTER III: Positive Historicism

3.1 Fischer and Ravizza

Positive historicist accounts that require an agent *has* a history of a certain sort have only, to my knowledge, been put forward by Fisher and Ravizza, and recently McKenna. Before offering general challenge to all positive historicist theories, as they are conventionally understood, I begin by examining the dialectic across these two accounts.

Positive historicism was first proposed by Fischer and Ravizza, largely in response to the intuitive pull of manipulations that nevertheless preserved non-historicist conditions, in an account that has become the “gold standard for cutting-edge defenses of compatibilism.”³⁸ However, it is important to recognize that the gold in this standard is constituted by their account of *reasons-responsiveness*, which refers to a structural feature of a person’s psychology at the time of action, and is thus non-historical. Since historical constraints are supplementary to and independent of non-historical constraints, in discussing the plausibility of *historicism* we should be willing to grant the compatibilist whatever he wishes for the non-historical part of his account. Thus, I leave aside whether Fischer and Ravizza’s account of reasons-responsiveness plays a part in the non-historical dimension of the most plausible account of moral responsibility, and we should assume going forward that in discussions of manipulation and historicism, all manipulated agents satisfy the necessary non-historicist conditions—what Pereboom calls the “causal integrationist conditions,”³⁹ McKenna calls the “compatibilist-friendly agential structure,”⁴⁰ and what Kane believes is preserved during what he calls “covert non-constraining manipulation.”⁴¹

The historical aspect of Fischer and Ravizza’s account comes from their ownership condition: the agent must have taken responsibility for his action, meaning that the mechanism that drives his

³⁸ McKenna and Coates (2015).

³⁹ Pereboom (1995), 100.

⁴⁰ McKenna (2012b), 149.

⁴¹ Kane (1996), 64-65.

action must be *his own*.⁴² Fischer and Ravizza cash out this common compatibilist criterion by delineating three conditions that must obtain:

1. The agent sees himself as an agent who is the source of certain causally effective choices.⁴³
2. The agent must see himself as an appropriate target for reactive attitudes.⁴⁴
3. An agent must arrive at (1) and (2) in an appropriate way.⁴⁵

This ownership condition contains both subjective and objective components, as well as non-historicist and historicist components. Fortunately, they align; (1) and (2) are both non-historical because they refer to a structural state of the agent, and subjective because they place constraints on how an agent subjectively relates to himself. Fischer and Ravizza recognize this. In fact their justification for (3) comes from the worry that an agent could be electronically-induced to be in the relevant subjective states described in (1) and (2).⁴⁶ I will not discuss the plausibility of (1) and (2) as they are non-historical and I have agreed to give the compatibilist whatever non-historicist account he likes—I also believe good challenges have been articulated.⁴⁷ The heavy historicist lifting is done by (3), which places (objective) constraints on how the agent *came to be* in these subjective psychological states. How do they explicate this crucial notion of appropriateness?

Earlier, we specified the third condition on taking responsibility as follows: the agent's view of himself must be based on his evidence in an appropriate way. Obviously this is abstract and schematic. This condition is intended (in part) to imply that an individual who has been electronically induced to have the relevant view of himself (and thus satisfy the first two conditions on taking responsibility) has *not* formed his view of himself in the appropriate way. But the relevant notion of appropriateness must remain unanalyzed.⁴⁸

For the only historical constraint in their account, this is anticlimactic. The challenge from manipulation arguments for incompatibilism (and therefore, against the soft-line historicist reply) is that there is no principled difference between a manipulative history and one found in a deterministic

⁴² Fischer and Ravizza (1998), 207.

⁴³ Ibid., 210.

⁴⁴ Ibid., 211.

⁴⁵ Ibid., 213.

⁴⁶ Ibid., 236.

⁴⁷ For objections, see Eshleman (2001), McKenna (2000), and Mele (2000).

⁴⁸ Fischer and Ravizza (1998), 236.

universe. Compatibilists disagree, as we expect, but for a compatibilist to simply assert that there are appropriate and inappropriate causal etiologies for non-historically identical states, and expect that to be persuasive, hardly pushes the dialectic forward. Sticking in an unspecified term to vindicate historicist intuitions by fiat leaves us with no principled basis for distinguishing appropriate and inappropriate histories other than the tug of intuitions about specific cases. Fortunately for Fischer and Ravizza, McKenna has recently put forward, though not explicitly defended, a positive historical constraint that does specify what an appropriate history for satisfying non-historical conditions is. It is this more refined account that we turn to next.

3.2 McKenna

McKenna argues for a positive historical account by way of an apparent deficiency with negative historical (NH) accounts, which he approximates as:⁴⁹

NH: An agent A-s freely and is morally responsible for doing so only if, with respect to the causal springs of her A-ing, she does not have a history that includes the acquisition of any unsheddable values through means that bypassed her ability to critically acquire, assess, and sustain them.

To highlight what he sees as a deficiency, McKenna gives us a variation of the original Ann/Beth manipulation case. He asks us to contrast two alternative pre-manipulation histories for Beth. Beth Active, from an early age, was obsessed with the life of mind and had a clear trajectory toward the same philosophically dominated-lifestyle possessed by Ann. But she took great effort in shaping her moral personality, and through constant reflection and determination, attenuated these academic proclivities and came to value other things. Beth Passive, on the other hand, was a sort of “value sponge” who just picked up whatever values she was exposed to, with no care for which values were more important or meaningful. Both get manipulated to become Ann’s value-twins with all her unsheddable values toward philosophy, and we’re expected to judge that neither Beth is responsible for the actions that ensue. This is the verdict the historicist wants. But McKenna believes NH gets the

⁴⁹ McKenna (2016), 95.

right answer for the wrong reason:⁵⁰

If we are to regard as freedom-and-responsibility undermining the intervention into Beth Passive's psychic life whereby very different unsheddable values are covertly forced upon her...it is not because Beth Passive was robbed of a moral personality that she came to possess under her own steam...Of course, being robbed of a moral personality in which one has invested is a rationale available for the case of Beth Active. [But] Beth Passive did not acquire her character so formed under her own steam; she was passive with respect to the acquisition of it in the history of her own life. Why then should such intervention be regarded as objectionable as it bears on Beth Passive's acting from a set of covertly and coercively installed unsheddable values? Unlike Beth Active, Beth Passive is not robbed of anything in which she has made an investment; having allowed the vagaries of life to buffet her about as they did, the acquisition of her values prior to the manipulation was just as happenstance as were the values that were later covertly installed in her through a process of manipulation.

To McKenna, a more compelling reason to account for Beth's non-responsibility is that she was robbed of the opportunity to fashion her own values. Despite her value-sponge existence pre-manipulation, McKenna still thinks that pre-manipulation Beth Passive was responsible, because "by omitting to engage in any active way, she allowed herself to arrive at the person that she came to be."⁵¹ Hence, Beth Passive was responsible in her pre-manipulation history since she had the *opportunity* (despite not taking it) to shape her own moral personality: "what is so important about an agent's having a history that lacks the acquisition of pertinent values through means bypassing her ability to critically assess them is that she thereby *has* a history that afforded her an opportunity to shape her moral personality for herself from abilities that she did possess."⁵² By contrast, in her post-manipulation phase, Beth Passive is stripped of her opportunity to assess her values since she becomes so enamored with philosophy, and hence she loses her responsibility. But if this reasoning is sound, then the underlying principle is positive, not negative.⁵³

PH: An agent performs a directly free act and is directly morally responsible for it only if any unsheddable values playing a role in the production of her action arose from a history whereby she was afforded the opportunity to critically assess, endorse, and sustain them from abilities that she possessed, and so none were acquired through means that bypassed those abilities.

⁵⁰ Ibid., 96.

⁵¹ Ibid., 97.

⁵² Ibid.

⁵³ Ibid.

As a first cut, this is certainly an improvement over Fischer and Ravizza's positive principle; it unpacks "appropriate" causal histories by requiring that autonomy-preserving ones have a history in which the agent was able to critically assess any unsheddable values she acts through (she had the opportunity), which seems like a plausible way of taking ownership. We can imagine Beth following a path like this: at some point when she is a morally responsible agent, her parents expose her to an academic way of life, which she blindly adopts. A few years later, her college roommates expose her to way of life that doesn't prioritize philosophy and which she also blindly adopts as a true value-sponge. In all these cases, Beth Passive has but does not act on an opportunity to assess new values, and so whenever she is exposed to a new lifestyle, the values associated with it which she acquires *are* unsheddable—but when she acts from these unsheddable values, she is nevertheless responsible.

However, there is a major complication with this use of the thought experiment: a *passive* moral personality *doesn't preclude the bypassing* of that personality. The first thing to note is that a negative historicist will not make sense of Beth Passive's pre-manipulation responsibility by invoking her opportunities for critical assessment—such positive constraints are clearly out of play. It will rather be by appealing to something like Odysseus-style cases: Odysseus wants to hear the Sirens, knows that if he hears the Sirens he will go to them and die, but doesn't want to die, so orders his men to tie him to the ship. While the Sirens sing, though his men seem to compel him to stay on the ship against his wishes, he is hardly *compelled* to do so since he endorses this "compulsion" in light of his higher-order preferences. Mele echoes this point when he suggests that even if one is manipulated to acquire an unsheddable attitude, if he arranges to be manipulated or even happens to strongly identify with the attitude for reasons unrelated to the manipulation, this is not sufficient for him acting unfreely.⁵⁴ Beth is in a similar position, so says the negative historicist, except that she does not *consciously* endorse what happens to her. She, in a sense, implicitly endorses being a value sponge, because her (passive)

⁵⁴ Mele (1995), 171.

evaluative framework which she autonomously possesses is such that pretty much all first-order value changes are authentic.

But we cannot equivocate between authentic changes to Beth Passive's first-order values, and potentially inauthentic changes to her higher-order *passive nature* that condones such first-order value changes; while Beth Passive in some sense endorses *being a value sponge with respect to other values*, she does not necessarily endorse being a value sponge *with respect to her being a value sponge*. The negative historicist can note that when Beth Passive is turned into some Ann value-twin, her passive nature is *replaced* with a philosophically-bent one that is no longer passive with respect to first-order value acquisition, and very plausibly suggest that this bypasses her evaluative framework and so is not an authentic change to it. When we distinguish these two different levels of passivity, we can see that it is false that the "acquisition of her values prior to the manipulation was just as happenstance as were the values that were later covertly installed in her through a process of manipulation." In the latter case not only are her first-order values changed, additionally the higher-order value that determines *how* she changes her first-order values is as well. Thus, what explains Beth Passive's post-manipulation non-responsibility is not that she is robbed of the opportunity to fashion her moral personality, but that she *is* robbed of her (passive) moral personality.⁵⁵

McKenna might object that we should understand Beth Passive as passive *both* with respect to first-order value acquisition *and* with respect to her being passive. But if this is the case, even when neuroscientists change her higher-order passivity values, she retains responsibility, per Mele's criterion that interventions that would otherwise be bypassing are not if the intervention is supported by the agent's pro-attitudes. Moreover, the negative historicist can note that in this scenario Beth Passive has been given the opportunity not only to assess her first-order values (an opportunity which she doesn't

⁵⁵ Also, if the university dean knew that Beth was passive with respect to first-order value acquisition, he could have realized that he only needed to expose her to Ann's values and she would blindly adopt them, freely—no outlandish neuroscientific brainwashing necessary.

take), but also to assess her passivity (another opportunity which she doesn't take). If that is true, then she did have an opportunity to reflect on whether she would endorse such an intervention, and under positive historicism she would also be free when it occurs.⁵⁶ The case of Beth Passive only supports positive historicism if Beth Passive both is not responsible after the manipulation *and* her critical capacities have not been bypassed (and thus, some other positive condition is needed to explain her non-responsibility). However, it is clear that once we distinguish passivity from bypassing, either the negative historicist can argue that her critical capacities have been bypassed, or if they have not, can argue that she is still responsible.

3.3 Childhood Values and First Free Acts

Even if the Beth Passive case does not support positive historicism, there is nevertheless another apparent advantage: the way it accounts for taking ownership. In so far as one believes that Frankfurtian ownership (or any non-historicist conception of 'taking ownership') is too thin, a positive account will almost be necessary to vindicate the condition. This is because negative accounts, in not requiring (instant) agents to have a past to be responsible, must invoke some 'default-responsible' ownership of one's values, and this doesn't seem to respect the 'taking' part of taking ownership. Before addressing our next criticism of positive historicism, we must clarify why McKenna stresses *opportunity*: positive historicism requires that one only have an *opportunity* to critically assess rather than require that one have actually critically assessed in order to allow for cases like Beth Passive to retain their responsibility, so long as they have the chance to assess their values, even if they do not take advantage of that opportunity.⁵⁷ Second, the process of critically assessing and endorsing one's values cannot itself be something that requires volitional control, otherwise the historicist would run into a

⁵⁶ After all, McKenna presumably wants to, like Mele, allow for autonomy preservation in Odysseus style cases, and PH could only do this if we understood the opportunity of rational reflection as including cases where an agent arranges for some change to his evaluative abilities, the completion of which removes from his evaluative abilities the pro-attitude of valuing that change in the first place.

⁵⁷ McKenna (2016), 97.

regress where any free act requires some prior free act (of assessing and endorsing) that would make it impossible for us to become free.⁵⁸ Thus, McKenna means for us to interpret “the opportunity to critically assess, endorse, and sustain from abilities that she possessed” in a way that only requires “a proper degree of cognitive control or activity.”⁵⁹

However, Haji questions whether this requirement is actually plausible. He starts with the noncontroversial observation that if we have moral responsibility at all, we do not have it from the instant we are born, but develop it over time. Hence, when talking about agent’s histories, it is instructive to specify whether some change occurs during their initial childhood and pre-agency stage, or after they become a morally responsible agent who has the relevant critical capacities for intentional and deliberate action.⁶⁰ He begins an apparent counterexample:⁶¹

Consider a scenario in which elder Beth Passive at some time at which, we assume, she is a free agent, A-s, an unsheddable value, W, is in A’s etiology, and she acquired W at some time during her childhood when her initial scheme—her “pre-moral personality”—was in embryo. At the time of W’s acquisition, it is false that little Beth Passive—rechristen her Beth Passive-A—was afforded the opportunity to assess critically, endorse, and sustain W because at this time she lacked the relevant abilities. Bear in mind that an agent, such as Beth Passive-A, can acquire W either during childhood or during the post-childhood free agency stage. If, as in our hypothetical scenario, she acquires unsheddable W at the pre-free agency stage, and W is a causal antecedent of some action, A, perhaps the first supposedly free action that she performs later at some time when she is a free agent, then A will not satisfy PH. So, Beth Passive-A will not be morally responsible for A-ing by the lights of PH.

Haji is expressing the idea that we are responsible when acting from values we acquired when we were very young. But if we acquired such values when we were very young, we acquired them long before we had the sort of cognitive control McKenna has in mind when requiring that unsheddable values arise in a history where an agent has an opportunity to exercise rational reflection. McKenna’s historicism would then have the implausible implication that we would not be responsible for many of the values we acquired in childhood (add genetic or innate unsheddable values). Since these do not

⁵⁸ Ibid., 99.

⁵⁹ Ibid.

⁶⁰ Haji’s account of the two will be more fleshed out when I discuss his negative historicist account.

⁶¹ Haji (2013), 199.

arise from a history where agents have the opportunity to critically assess them, if we interpret McKenna's account in that way, the grievance has bite. However, it is not clear that this is what McKenna means, and moreover we are interested in whether the strongest version of positive historicism is plausible, so we next turn to whether we can clarify or revise the account in light of this. To do so, we need to be extremely careful with how we invoke unsheddability.

3.4 Unsheddability, Reconsidered

Recall that an unsheddable value (shorthand for a practically unsheddable value) is one that an agent cannot, for all intents and purposes, eradicate or attenuate *over some period of time*. Unhelpfully, neither McKenna (nor Haji, though I will discuss his account in the following chapter) specify the time interval they have in mind in many of their thought experiments, which is a major source of ambiguity. Let us fix the time intervals for unsheddable values:

O-Unsheddability. Let an O-unsheddable value for action A be a value in A's etiology that an agent cannot eradicate or attenuate over the entire course of the value's possession (from the moment of the value's acquisition to the completion of A).

X-Unsheddability. Let an X-unsheddable value for action A be a value in A's etiology that an agent cannot eradicate or attenuate over the course of action A.

And with their contrastive, collectively exhaustive counterparts:

O-Sheddability. Let an O-sheddable value for action A be a value in A's etiology that an agent can eradicate or attenuate over the entire course of the value's possession (from the moment of the value's acquisition to the completion of A).

X-Sheddability. Let an X-sheddable value for action A be a value in A's etiology that an agent can eradicate or attenuate over the course of action A.

Clearly, if a value is unsheddable over some time interval, then it is unsheddable over some interval within that time interval; a value being O-unsheddable entails it being X-unsheddable.⁶² Conversely, a

⁶² Inversely, a value being O-unsheddable over some small interval of time might actually not be O-unsheddable over some larger interval of time, as an agent develops. This is important because even if someone is turned into the value twin of someone else, his qualitatively identical values are not automatically equally sheddable. Consider Bill, who spends his entire life reflecting on altruism and through habit developing a propensity to donate that is X-unsheddable when he acts from it, though it is O-sheddable over the duration of having the value. If Bob, who is manipulated to be Bill's value twin, donates from a qualitatively identical X-unsheddable value, it is also O-unsheddable because he acquired that value much later than Bill did.

value being X-unsheddable does not tell us whether it is also O-unsheddable. An agent may not be able to mitigate the value at will over the course of the action, but he may have been able to in the weeks or months leading up to that action. Thus, his X-unsheddable values might actually be O-sheddable. On the other hand, a value that is O-sheddable does not entail that it is X-sheddable, while a value that is X-sheddable entails its being O-sheddable. Which of these does McKenna invoke? He gives a couple unequivocal hints.⁶³

1. This assumption is licensed by the highly plausible empirical presupposition that most actual persons have and act from some range of values that *they are not able to alter at the time proximally coincident with their acting*; their values, or at any rate most of them, are in this sense relatively stable, fixed features of their psyches, perhaps subject to revision over larger stretches of time, but *not simply alterable on the fly as the contingencies of practical life present themselves*.⁶⁴
2. I suppose some of the values a person possesses at a time are ones that it is up to her whether she possesses them at that time; were it so, she would be able to lose or substantially downgrade them *at will*. But I strongly suspect that most values that most actual persons possess at a time are such that, *at that very time, those values are practically unsheddable for them*...at that time their values are ones which are such that, in normal practical contexts, it is simply not up to them whether they possess those values, nor is it up to them the relative strength those values have from them.⁶⁵

Thus, McKenna is using unsheddability to refer to X-unsheddable values that an agent cannot attenuate at the time (or over the course) of his action—he cannot attenuate them at will. This usage of unsheddability is different (and of a much shorter time interval) than what Mele had in mind when he originally stipulated the technical term, and Mele makes this very complaint about McKenna’s usage.⁶⁶ While a taxonomy of different types of unsheddability is not an end in itself, this verbal difference does mean we should be careful with what we mean when we invoke unsheddability, and what we intend to use it for.

Is invoking X-unsheddable values the most useful convention when discussing historicism? I do not believe it is. First, as Mele note in his complaint about McKenna’s usage, “in actual people, most values are [X]-unsheddable...In fact, ‘most’ is an understatement. (How frequently do we

⁶³ Italics mine.

⁶⁴ McKenna (2016), 88-89.

⁶⁵ McKenna (2012a), 148.

⁶⁶ See Mele (2013), 157-158 for a similar observation on how McKenna uses unsheddability.

eradicate beliefs at will or significantly reduce or increase the strength of a desire of ours at will?).”⁶⁷

If basically *all* the values we act on are stuck to us in the moment of action, X-unsheddability cannot be a useful convention. Consider manipulators that electronically induce some agent to have a desire to steal which is X-unsheddable months later when he steals, with that value in his action’s causal etiology. Is the agent responsible for stealing? It depends how we fill in the details. Either the X-unsheddable value was also unsheddable over some longer time interval or it was not. If it *was*, then the manipulators injected a value that the agent had no chance of eradicating over a much longer history, and this seems like a prime example of manipulation bypassing one’s critical capacities and undermining his responsibility. But if it *wasn’t*, then over the course of having it, he *was able to eradicate or attenuate* its strength. Harkening back to Mele’s earlier point that being able to shed a value does seem like plausible grounds for it being authentic (if one does sustain it), then we freely act from many of our X-unsheddable values that are sheddable over some longer period. And given that basically all of our values that we act from are X-unsheddable and that the historicist wants to argue that we are (often) morally responsible for our actions, this is exactly the result he would like.

There is a further question. Suppose Mele’s reasoning is sound, and a value’s being sheddable over some time period longer than the moments it is causally implicated in some action provides, along with other compatibilist criteria, good reason for thinking it is autonomously possessed. Does this apply all the way back to the moment the value was acquired? In other words, if a value is O-sheddable, and over the entire course of the value’s history the agent *did* have the ability to eradicate it or significantly attenuate its strength, does this give us good reasons for thinking he autonomously possesses it? On the one hand, it is hard to see why Mele’s reasoning, if correct, wouldn’t also apply here; an agent has a value that he really could have eradicated but has nevertheless sustained, and thus he seems to have the sort of ability that constitutes autonomy. On the other hand, Mele himself, when

⁶⁷ Ibid, 158.

he does specify the time interval of unsheddability, always discusses it over the time span of weeks or months. Since values can be unsheddable over months but sheddable over the span of their entire duration of possession, he does not indicate whether he sees his principle applying over this longer period of time. And it does seem intuitively plausible that engineered values that are unsheddable over very long periods of time are not autonomously possessed, even if they technically could have been shed over the entirety of their duration.

I will not argue here for the correct time interval for sheddability being sufficient for autonomy, and *a fortiori* whether a value being O-sheddable is sufficient for it being autonomously possessed (for the same reason as in Mele's argument that X-sheddable values are autonomously possessed). While the answer *does* affect the ultimate degree of strength of my argument against positive historicism, it does not affect the overarching structure (or, in the worst possible case, weaken it enough to remove its force). And that is because there seems to be widespread acceptance that any account of moral responsibility that is not implausibly revisionary will uphold that at least some of our O-unsheddable values—like innate and early childhood values that we acquired long before we had the ability to critically assess them—are autonomously possessed. As we have seen with Haji, and as we will encounter more later, the very fact that an account is perceived to be unable to account for the authenticity of such values is itself taken as a severe objection to its plausibility. My eventual argument against positive historicism (as McKenna understands it) is that it cannot account for the authenticity of O-unsheddable values. If it turns out that the same objection will actually apply to the authenticity of values that are unsheddable over shorter periods of time but are still O-sheddable, then the account will be even more implausibly revisionary.

3.5 McKenna and Positive Historicism, Defended

Let us return to McKenna's account, using the conception of unsheddability he endorses:

PH: An agent performs a directly free act and is directly morally responsible for it only if any X-unsheddable values playing a role in the production of her action arose from a history whereby she

was afforded the opportunity to critically assess, endorse, and sustain them from abilities that she possessed, and so none were acquired through means that bypassed those abilities.

We need to make a revision to PH in light of our discussion of unsheddability. Recall that a value being X-unsheddable is consistent with it also being either O-unsheddable or O-sheddable. A value can be X-unsheddable *either* because it in some sense passed through one's evaluative scheme and was increased in strength instead of eradicated, *or* because it bypassed one's evaluative scheme from the moment of acquisition to the moment of the action in which it played an unsheddable role. That is, either the X-unsheddable value is unsheddable over the course of the action, but as you extend the time interval farther back toward the value's acquisition, at some point over that longer time horizon it is sheddable (it is X-unsheddable but O-sheddable); or it remains unsheddable even as the time horizon is lengthened to the moment of its inception (it is O-unsheddable). PH, then, entails:

PH: An agent performs a directly free act and is directly morally responsible for it only if *either* any O-unsheddable values *or* any O-sheddable and X-unsheddable values playing a role in the production of her action arose from a history whereby she was afforded the opportunity to critically assess, endorse, and sustain them from abilities that she possessed, and so none were acquired through means that bypassed those abilities.

For reasons we have already discussed, we can ignore the component that requires that values that are O-sheddable and X-unsheddable be subject to rational reflection; historicists seem to agree that one freely acts from values that he could have attenuated at some point from when they were acquired to when he acted on them (O-sheddability) but nevertheless sustained. But even if they don't, I am happy to concede this to the positive historicist—it is nevertheless important to account for the authenticity of O-unsheddable values.⁶⁸ Thus, the plausibility of McKenna's positive historicism reduces to the plausibility of the following:

PH: An agent performs a directly free act and is directly morally responsible for it only if any O-unsheddable values playing a role in the production of her action arose from a history whereby she was afforded the opportunity to critically assess, endorse, and sustain them from abilities that she

⁶⁸ Moreover, it is plausible that a value *being* sheddable means that it *has* had the opportunity to be subject to rational reflection. While sheddability is a term of art stipulated by Mele and rational reflection, as used here, is one by McKenna, there is an open question of how they relate or should relate, and this seems like one plausible link. However, I will not explore this issue here.

possessed, and so none were acquired through means that bypassed those abilities.

This helps us better understand how to interpret “arose” in the construction. It is tempting to understand “arose” as meaning something like having the opportunity to “become unsheddable through a certain type of history.” However, if a value that plays a role in producing an action is said to *become* unsheddable after being critically assessed, endorsed, and sustained, this has to be referring to an unsheddable value that is O-sheddable (some type of unsheddability whose time interval is shorter than the interval from the moment of the value’s acquisition). An O-unsheddable value does not *become* unsheddable—it always is. This by no means suggests that rational reflection is not the sort of capacity that can make values unsheddable. Of course it can; it just can’t do this for O-unsheddable values. Hence, if the same ability, rational reflection, is used to account for the authenticity of O-unsheddable values, “arose from” really means something more like “existed through or persisted through.” In this case, even though rational reflection is not what crystallizes some value into the unsheddable one that plays the role in some action, one may still act freely from an O-unsheddable value given that he relates to it through rational reflection.

At least at first, this actually defuses Haji’s remaining criticism—that positive historicism cannot account for early childhood-acquired (or innate) unsheddable values because it cannot account for unsheddable values acquired before one develops rational reflection. Recall that Haji’s basic point is that, given we are not born with the critical capacities for rational reflection, if Beth Passive-A acquires an unsheddable value before such capacities develop it is false that “she was afforded the opportunity to assess critically, endorse, and sustain W because at this time she lacked the relevant abilities,” while Haji believes that Beth Passive-A could be responsible for acting from such a value. Haji contrasts Beth Passive-A with Beth Passive-B, who acquires a value “that is not unsheddable when first acquired during her pre-free agency stage in her life” but which “crystallizes into an unsheddable value at a time when actions that express it may satisfy PH” and where Beth Passive B

does not “actively contribute” to it becoming unsheddable.⁶⁹ To Haji, we would seem to appeal to an arbitrary difference if we ascribe asymmetric responsibility to the two Beth’s who passively take in their values, but one of whom has a value become unsheddable just after she becomes a free agent, and the other just before.

We need to perform a little surgery on the thought experiment in light of the previous discussion of sheddability. First, we should note that Haji is, like McKenna, understanding unsheddability in terms other than O-unsheddability—this is clear because he talks of a value that starts off as sheddable “hardening” into an unsheddable value. But, if Beth Passive-B has a value become unsheddable, whereas Beth Passive-A has a value that has been unsheddable since its inception, one is O-unsheddable and the other is O-sheddable, and this is exactly the relevant difference between the two cases which could explain asymmetric responsibility ascriptions. I take it that Haji asks us to only focus on the scenarios in which she does not “actively contribute” to the value becoming unsheddable to reinforce the idea that it is arbitrary that both Beth Passives do not do anything that result in the value’s being unsheddable—it is just the arbitrary difference in time. Let us imagine unsheddability results from a random reconfiguration of her brain, and Beth Passive’s evaluative scheme is passive with respect to *any* first-order value acquisition, in order for her to retain autonomy as Haji wants. But if the value just one day becomes unsheddable in this way, it is just like the value is acquired as unsheddable—it being previously sheddable has nothing to do with any of its current properties relevant to responsibility, and so it essentially “renews” its time point of acquisition. The essential feature of Haji’s criticism, then, is that apparently on McKenna’s account if a person is given an O-unsheddable value a millionth of a second before he becomes the sort of agent with the capacity for critical reflection, then he will not be responsible when he later acts on it; but if he is given the same O-unsheddable value a millionth of a second after he becomes capable of rational reflection,

⁶⁹ Haji (2013), 199.

he will be later responsible for it—and this huge difference in responsibility ascriptions seems arbitrary.

If this implication were true of McKenna's account, I agree that it would be a powerful objection. But the objection only has bite if we use a now discredited conception of how we are supposed to understand "arose." The objection works if we understand McKenna's account as requiring an agent to have an opportunity to critically assess unsheddable values *as they arise*. In this case, if one lacks the requisite capacities, then clearly one cannot be free for actions stemming from values that arose before such capacities exist. However, as we have reviewed, if we require that one's critical capacities have a chance of making the value unsheddable when it arises, this is *not* a discussion of O-unsheddability, and so we cannot understand "arose" in this way in this context. Rather, as we have said, the most charitable version of McKenna's account requires an agent to have an opportunity to critically assess O-unsheddable values *once they have arisen*. But since this is the case, it *does not matter whether the requisite capacities are around* when the unsheddable value pops into a person's evaluative scheme; the positive historical requirement is satisfied as long as at some point the agent has a history that gives him the opportunity to critically assess values that are already unsheddably inside his evaluative framework. There need be, then, no difference in our responsibility ascriptions for someone who is given an O-unsheddable value when he already has critical capacities for rational reflection and who immediately assesses it, compared to someone who is given the same value prior to having such capacities and only later subjects it to rational reflection. Indeed, if Haji's objection ruled out unsheddable values that were acquired pre-agency since they did not *become* unsheddable through rational reflection, this would also rule out unsheddable values acquired post-agency that did not *become* unsheddable through rational reflection—but then this is just a restatement of the thesis that one can only be free for acting from O-sheddable values. And since Haji interprets McKenna's account as allowing agents to be morally responsible for acting from O-unsheddable values acquired post-agency,

his objection of arbitrariness does not, as stated, threaten McKenna's account.

3.6 Optimism Toward Positive Historicism, Reconsidered

Clarifying Haji's criticism with a careful eye toward unsheddability shows us how the most charitable understanding of McKenna's positive historicism can deflate it. But it also reveals a major shortcoming of McKenna's view. The positive historicist aspect comes from one having the opportunity to undergo rational reflection, which can critically assess and endorse O-unsheddable values, and which is the ability that engineered O-unsheddable values cannot bypass without undermining responsibility. How exactly must rational reflection relate to any given O-unsheddable value? Either it is the sort of thing that can affect whether the value is O-unsheddable, or it can't. And we know that rational reflection cannot affect whether the value is O-unsheddable, by definition.

Thus, a logical constraint on what rational reflection is, *when it relates to O-unsheddable values*, is that it cannot be the sort of capacity that can actually significantly attenuate a value over the course of its history. But if this is true then, as I will now argue, positive historicism seems no more plausible than a comparable non-historicist position. To see why, we should first see what McKenna says to justify the positive historical aspects of his account over some non-historicist alternative:⁷⁰

- What is so important about an agent's having a history that lacks the acquisition of pertinent values through means bypassing her ability to critically assess them is that she thereby has a history that afforded her an opportunity to shape her moral personality for herself from abilities that she did possess.
- What is required with respect to the causal springs of a free act is that an agent had a history in which she had the opportunity to fashion her evaluative standpoint for herself from abilities that she possessed.
- [We must emphasize] the role of having a history—one that makes room for the opportunities whereby one is able to shape her moral character.
- A requirement of acting from an unsheddable value, as PH suggests, is that an agent had a history which provided conditions enabling her to engage in some relatively complex evaluations regarding, first, what is valuable, and, second, how in light of this she would judge it best to structure her practical affairs.

The first thing to remember is that this process of shaping one's moral personality from pre-existing

⁷⁰ All quoted from McKenna (2016), 97.

abilities cannot itself be one that requires free action; if every free action derives from or requires some prior free act (of critically assessing some value), we run into an infinite regress that makes it impossible for responsibility to get off the ground. McKenna is acutely aware of this, noting that rational reflection should not be analyzed in terms of action at all, but rather some type of “cognitive control.”

But if the fact that some prior shaping was *freely* conducted cannot be what is driving our intuitions that shaping is an important historical requirement on responsibility, what does? It seems that the most plausible, and perhaps only, explanation implicit in McKenna’s justifications is that the process of shaping, even if not a process which is done freely, still *could make a difference* for how one’s moral character turns out. The language of “shaping” and “fashioning for oneself” are extremely suggestive of the ability to control the diverging paths one’s moral character might take. Of course, since we are talking about all of this in the context of determinism, this is nothing like the ability to make the actual world proceed other than it actually does, but rather to preserve some sense in which one’s character, through his own efforts, is within his own control. That this conception of shaping and fashioning one’s moral personality is what is lending positive historicism its intuitive support seems clear when McKenna applies PH to explain why Suzie Instant might actually not be free: “PH helps to call forth the concern that there is something objectionable about regarding Suzie Instant as free and responsible for her acts just after being brought into existence. Her evaluative framework was fully fixed for her without her so much as having a chance to have engaged in the shaping of it.”⁷¹ An evaluative framework that is fully fixed is one the strength of whose component features one lacks the requisite control over.

But if the plausibility of positive historicism rests on the intuition that one can shape, not freely, but in this sense, his moral personality, then the plausibility of positive historicism rests on *an*

⁷¹ Ibid., 98.

*ability that rational reflection does not have with respect to O-unsheddable values.*⁷² Since McKenna's account must allow rational reflection to apply to O-unsheddable values, rational reflection must apply to values for which it is practically impossible over the entire duration of having the value for an agent to significantly attenuate its strength. But in what sense can one "shape" a value whose strength, over the entire course of possessing it, he cannot alter? When an O-unsheddable value passes through one's cognitive sieve in whatever way rational reflection requires, rational reflection has no way of genuinely "shaping" that value. This makes it difficult to see why positive historicism has any upshot over a comparable non-historicist account. If positive historicism requires that one has a history in which, with the requisite complex critical capacities, one examines and critically reflects on one's values, even if those examinations and critical reflections have no chance of making a difference for an agent actually possessing those values, this seems no more plausible an account of responsibility than one merely identifying in some particular way with an action at the time he engages in it. It would seem unprincipled to absolve someone of blame for committing a murder with which he identified on the grounds that he hadn't really reflected on his murderous values, but then to blame a qualitatively identical murderer who had really really thought about those values on the grounds that he had the chance to really consider having murderous intentions even when such prolonged reflection had no chance of making a difference to having such malicious values. If one disregards non-historicism because he doesn't believe that identification is an adequate notion of taking responsibility, it is unclear why identification plus ineffectual opportunities would be.

If this is true, it helps explicate why it is crucial to specify the duration of unsheddability when the concept is being invoked. McKenna's original account is ambiguous between O-unsheddable X-

⁷² To be clear, this is just how rational reflection relates to O-unsheddable values. Of course rational reflection is the sort of ability that can shape our moral personality and affect whether values are sheddable over shorter intervals of time; recall the earlier example of the man who reflects on his altruism and through habit builds up what becomes an X-unsheddable (and even, perhaps, unsheddable over the span of months or even years) value to donate when he does.

unsheddable values and O-sheddable X-unsheddable values. O-sheddable X-unsheddable values are ones for which one *can*, through rational reflection, actually change whether such values are ultimately fixed in his evaluative framework at the time of action, and in this sense *make a difference* for the person one becomes. O-unsheddable X-unsheddable values are ones for which, even when subjected to rational reflection, one cannot. If the plausibility of the account hinges on what we think of when we think of how rational reflection applies to O-sheddable X-unsheddable values, we get an intuition bait and switch where we're not thinking about what we need to be considering.

While this applies most clearly to McKenna's account, it spells out a more global worry with *any* positive historicist account. If we abstract away from McKenna's specific proposal for what relation the agent needs to have during his history in order to be responsible—the opportunity for rational reflection—in its thinnest form, positive historicism, it appears, just requires that an agent have a history in which he relates to his O-unsheddable values in a particular way (and it is in relating in this way that he takes ownership over those values, ensuring they have not bypassed him).

PH (general structure): An agent performs a directly free act and is directly morally responsible for it only if for any O-unsheddable values playing a role in the production of her action, she had a history in which she had (or had the opportunity to have) some critical relation to them.

Again, for any positive historicist account whose structure matches the one just described, the critical capacity function can relate to one's values in one of two ways: either it can make a difference to the strength of those values, or it cannot make a difference. Because of how O-unsheddability is stipulated, it cannot make a difference, and since it can't, then the critical relation doesn't actually make a difference for the person one becomes. One cannot genuinely shape a moral personality with respect to a class of values some of which all compatibilists believe are authentic, and the account seems no more plausible than the same critical relation in a non-historical form.⁷³

⁷³ The positive historicist may argue that there are no O-unsheddable values (e.g. while some values might always have to be fixed, in principle you can carry out rational reflection on one part of your evaluative scheme from the perspective of another part of that scheme), which if true would defuse the challenge. I will not further explore this line of argument.

At this point one might worry that we have stacked the cards against historicism: if there's no relation an agent could have (or could have the opportunity to have) with respect to values for which he has never been practically able to shed, this seems to suggest that we're committed to non-historicism if we want to retain any conception of moral responsibility when acting from O-unsheddable values. But this would be moving too quickly: even if there's no way an agent *could relate*, this only makes a strike against a *positive* historical account. Negative historical accounts do not require an agent to have done *anything* to be responsible for O-unsheddable values—only that certain events have *not occurred*. It is the plausibility of this class of historical accounts to which we now turn.

CHAPTER IV: Negative Historicism

4.1 Mele

Negative historicist accounts have been most prominently defended by Mele, and Haji and Cuypers, whose accounts I focus on in this chapter.⁷⁴ Mele, the originator of the Ann/Beth manipulation cases, believes that what best explains the asymmetric contrast is not that responsible agents must have certain pasts, but rather that they must lack certain ones. He is clear about defending a negative historical account, which “requires [not] that the agent have a history of a certain kind, but rather that he lack a certain kind of history—a history yielding what I have called ‘compulsion’ of [that attitude].”⁷⁵ For Mele, one may authentically possess some pro-attitude (from which he can autonomously and freely act) only if he is not compelled to have that attitude (unless he has arranged for himself to be compelled to have that attitude). To approximate sufficient conditions for compulsion, he writes:⁷⁶

If an agent S comes to possess a pro-attitude P in a way that [i] bypasses S’s (perhaps relatively modest) capacities for control over his mental life; and the bypassing issues in [ii] S’s being practically unable to shed P; and the bypassing was [iii] not itself arranged (or performed) by S; and [iv] S neither presently possesses nor earlier possessed pro-attitudes that would support his identifying with P, with the exception of pro-attitudes that are themselves practically unsheddable products of unsolicited bypassing; then S is compelled* to possess P.

Mele’s account helps underscore the difference between positive and negative accounts. While both invoke some conception of bypassing to distinguish autonomy and heteronomy, a positive account requires an agent to have gone through some measures to ensure that their critical capacities have not been bypassed, whereas a negative account requires an agent have *not* gone through some

⁷⁴ To my knowledge, the first, and only other, person to propose a negative historicist account was John Christmas (1991), with three conditions:

- (i) A person P is autonomous relative to some desire D if it is the case that P did not resist the development of D when attending to this process of development, or P would not have resisted that development had P attended to the process.
- (ii) The lack of resistance to the development of D did not take place (or would not have) under the influence of factors that inhibit self-reflection.
- (iii) The self-reflection involved in condition (i) [sic] is (minimally) rational and involves no self-deception.

However, I will not focus on Christmas’ account here, since I believe Mele (1993) provides a compelling critique, and indeed builds his own account to accentuate the merits of Christmas’ without its complications.

⁷⁵ Mele (1995), 172.

⁷⁶ Ibid.

circumstances, which ensures their critical capacities have not been bypassed. Perhaps more intuitively, a positive account assumes that some non-historical set of control properties is default non-autonomous, and specifies histories that result in autonomy; a negative account assumes that some non-historical set of control properties is default autonomous, further specifying histories that subvert autonomy. Later we will explore how Mele understands unsheddability in light of our previous discussion of time-stamping it. For now we can simply note that [i] and [ii] corroborate the intuition that *unsheddable* values are particularly worrisome when they are implanted, while [iii] and [iv] anticipate our discussion of Odysseus-style cases.⁷⁷ It is crucial to note, however, that Mele only specifies a certain set of *sufficient* conditions for compulsion. This is at best a partial account of responsibility, since we are left without *necessary* conditions for responsibility or non-responsibility that, together with sufficient ones, allow us to move from any complete description of a manipulation case to a verdict on whether responsibility has been subverted.

Nevertheless, we can appreciate some advantages of negative historicism. As with positive historicism, it allows us to say that Beth is not responsible, but for a slightly different reason: *not* for the fact that she is robbed of the opportunity to fashion for herself a moral personality, but simply because she is robbed of her moral personality. This difference is crucial in allowing negative historicism to deliver the intuitive verdict that instant agents like Suzie Instant are responsible, while retaining the intuitive judgment that Beth is not free. McKenna believes that that there is a good case for treating Suzie Instant and Beth symmetrically:⁷⁸

[S]olely by virtue of nonhistorical properties just subsequent to her A-ing, [Suzie Instant] satisfies sufficient conditions for A-ing freely and being responsible for it. But, by hypothesis, Beth satisfies the same

⁷⁷ Mele (1995, 171) gives the example: suppose Beth wants to maximize her happiness, and that value is unrelated to the manipulation. If she is manipulated to then see philosophy as maximizing her happiness, and so she values philosophy for that reason, she plausibly is still free; if an implanted value is unsheddable for reasons not explained by the manipulation, then it may still be authentic. To give a less sci-fi example, suppose I go to a philosophy professor's lecture and become convinced that AI research is the best way to help the world. If I want to do the best thing to help the world, I now value doing AI research for that reason. I could hardly be said to have been manipulated (in a responsibility-undermining way) to value AI research. More accurately, it *resonated* with reasons I already had.

⁷⁸ McKenna (2012a), 271.

nonhistorical properties as does Suzie Instant when Beth A-s. Hence, if Suzie Instant A-s freely and is morally responsible for doing so, then Beth A-s freely and is also morally responsible for doing so. If, on the other hand, Beth does not A freely and is not morally responsible, it seems that the same ruling applies to Suzie Instant, since the case of Beth shows that the nonhistorical properties both satisfy are insufficient for acting freely.

McKenna suggests that we face a dilemma in adjudicating historicist and non-historicist theories: while we may intuitively think Beth is not responsible and Suzie Instant is, we seem to have reason for treating them symmetrically. Hence, we must either take the non-historicist theory and try to explain away our intuition that Beth is not responsible, or take the historicist route and try to explain away our intuition that Suzie Instant is. I have a difficult time seeing exactly why McKenna believes there is a reason for treating Beth and instant agents symmetrically, but the argument seems to only have bite if one presupposes a positive historicist account of moral responsibility. If one must *have* a history of a certain sort to be responsible, then clearly Beth and instant agents lack such a history. But if one must only *lack* a history of a certain sort to be responsible, then even while Beth and instant agents may be non-historical duplicates, literally not having a history is different from being radically reversed from your previous self. If the types of histories that negative historicism asserts as freedom-undermining include those that involve radical reversals which is not in the (non-existent) history of instantaneous creation, then negative historicism *does* have grounds for delivering an asymmetric judgment to Beth's (unfree) actions and instant agents' (free) actions.

4.2 Innate and Early Childhood Values

While negative historicism seems capable of delivering asymmetric responsibility verdicts about Beth in contrast with instant agents, there seems to be a problem lurking. I previously reviewed Haji's criticism that positive historicism cannot account for unsheddable values acquired before one has the capacity for rational reflection, and argued that the criticism, upon further consideration, does not have traction. Cuypers makes the same criticism of Mele's negative account, and when applied in this way, the challenge seems to have significantly more force. If, on Mele's account, one is compelled to

have an unsheddable value that is implanted in a way that bypasses his capacities, then it turns out that

Most of an agent's pro-attitudes are compelled (or inauthentic) and thus nonautonomous. The reason is that in the formation of the bulk of S's desires based on innate needs and socially acquired preferences, S's control-capacities are equally bypassed...I cannot, for example, change the strength of my strong desire for coq-au-vin and I am unable to eradicate my strong aversion to chicken-liver. Also, I cannot give up my preference for classical baroque-music; and I am incapable of fostering a preference for operetta-music to win the love of my mother-in-law. Neither the origin nor the genesis or formation-process of the greater part of my pro-attitudes (and anti-attitudes) falls under my control-capacities.⁷⁹

If we are compelled to have a value when it bypasses some control capacities, then since we are not born into the world kicking and screaming and with sophisticated capacities for control, there is a time period at which any unsheddable value we acquire is compelled—our control capacities are necessarily bypassed since they don't yet exist.

Mele's only direct reply to Cuypers is that Cuypers claims Mele to be committed to the principle that an agent can autonomously possess a pro-attitude only if its generation lacks control-bypassing. However, Mele is careful to draw a distinction between a value one is compelled to *acquire* and a value one is compelled to *possess*.⁸⁰

I am not suggesting that compelling a person to acquire a collection of values, as in the brainwashing scenarios, is *sufficient* for rendering a person nonautonomous. Whether someone's having been subjected to effective manipulation deprives him of autonomy depends upon whether those effects remain in force. A distinction is in order, then, between being compelled to acquire a value at a time and being compelled *to possess* a value over a stretch of time. I have been supposing that in the brainwashed Beth scenarios, she is no more likely to change her new values than are a mature Ann or [another similarly manipulated agent] to change their deeply entrenched values.

Mele suggests that there are two ways we can understand the Beth scenario. In one scenario, Beth is implanted with unsheddable values, but the effects of the manipulation do not “remain in force” and she is then a valid candidate for responsibility. In the other, she is implanted with unsheddable values, and the degree to which she is stuck with them remains the same, in which case she is nonautonomous. Hence, even if one is compelled to acquire a value—it is unsheddable at the moment it is implanted—

⁷⁹ Cuypers (2006), 182.

⁸⁰ Mele (1995), 158-159.

its unsheddability might turn into sheddability over a longer stretch of time. This is exactly what we'd expect; a longer period of time gives one a greater opportunity to reflect and attenuate or sustain a value he might have originally been caught up in the throes of. Expressing this in our time-stamping vernacular: an implanted value that one immediately acts on might be O-unsheddable (say, in the two hour time period between when the value is implanted and when one acts on it first), but it might be O-sheddable when one acts on it later (when there is a two week period between when it was implanted and when one acts on it).

Does the application of this distinction defuse Cuyper's objection? In one sense, yes, but in a larger sense, no. Many of our innate and socially acquired preferences (which were acquired before we had control capacities) might be O-unsheddable when we first act on them, but actually O-sheddable over the course of many years. If they are then X-sheddable or even sheddable over some two-week time-stamp arbitrarily later in our lives, we autonomously act on them. Thus, not all of the innate or socially acquired preferences we have from our pre-control period of our life are, by Mele's account, necessarily inauthentic. However, when Cuypers says he cannot give up his preference for classical baroque, we can plausibly interpret him as saying that he cannot now nor has he ever been able to give it up, yet he does not see it as constraining his autonomy. For all he cares, decades later the value remains O-unsheddable—it was compelled when he acquired it, and since the effects remain in full force, he is still compelled, in Mele's sense, to possess it—yet this does not subvert his responsibility when acting from it. This is just a restatement of the common sense belief that we can be responsible for acting from some O-unsheddable values—even those that were acquired very early in our life. Thus, Mele's account cannot explicate the authenticity of O-unsheddable values that were acquired before one had control capacities; in so far as such values can be authentic, the account is deficient.

4.3 Haji and Cuypers

Haji and Cuypers propose a negative historicist account that builds off the strengths of Mele's while

retaining an ability to account for innate and early childhood acquired values.⁸¹ They begin by distinguishing two stages in the development of moral agency: before someone becomes a free agent, and after. In order for a person to be capable of the sort of deliberative, intentional action needed for moral responsibility, Haji and Cuypers note that one must have some psychological basis for evaluative reasoning, what they call an “evaluative scheme.”⁸² This contains components such as:

- i. Normative standards for assessing reasons, and how an agent goes about deciding.
- ii. Desires, beliefs, and plans to attain long term goals.
- iii. Deliberative principles to adjudicate practical reasoning.
- iv. Motivation to act on the basis of i-iii.

The pre-agency stage corresponds to the point before a person has a (sophisticated enough) evaluative scheme to warrant free action; once this develops, the person enters the post-agency stage with their *initial evaluative scheme*. This contains all of the values that the person acquired innately or in early childhood (as proto-values). In assessing whether such an initial evaluative scheme is authentic such that the person is responsible for his actions issuing from it, Haji and Cuypers suggest that initial schemes can only be “relationally authentic”: they are authentic only insofar as they do not subvert responsibility for later behavior.⁸³ This is equivalent to saying that there is no *historical* requirement on the authenticity of initial evaluative schemes; they are only inauthentic in so far as they contain elements that are inauthentic for non-historical reasons (such as containing an irresistible desire).

Of course, one’s psychological basis for evaluative reasoning changes over time. An *evolved scheme* is the evaluative scheme that results from changes to the initial scheme. Haji and Cuypers argue that an evolved scheme is authentic only if it is “causally related to initial schemes through alterations

⁸¹ Haji and Cuypers (2007, 343) insist that their view is a “hybrid” between historicism and non-historicism because it does not require of agents “that one have a past but it does require that one not have certain kinds of past.” However, this fits unproblematically into the negative historicist position I have sketched out. More recently, Haji (2013, 185) defends the same account, noting that it is a “hybrid” since “it implies that responsibility does not require that one have a past, but should one have a past, it must be a past of a certain sort.” That this new way of presenting the same account has a positive ring shows why it is important to cash out the positive/negative distinction in event terms (as it turns out that the past one must have is one that lacks various events).

⁸² Haji and Cuypers (2007), 350.

⁸³ Ibid., 359.

that can only occur through an agent's exercise of her own deliberative control."⁸⁴ That is, one can only freely act from an evolved scheme if the changes to one's initial scheme that resulted in that evolved scheme were facilitated *through* that initial scheme (or, later on, if the changes to one's previous evolved scheme were facilitated through that previous evolved scheme, resulting in a new evolved evaluative scheme). Perhaps more intuitively and most generally this means that every evaluative scheme implies some conceptual list of acceptable (authentic) modifications to that scheme (though the common denominator of such changes is that they occur through the agent's deliberative control), and evolved schemes must be connected to one another such that each new scheme was on the authentic list of changes that could be done to the previous scheme, according to the previously scheme. Hence, Haji and Cuyper's account essentially boils down to two authenticity conditions:

1. An initial evaluative scheme is (responsibility-wise) authentic if it fits the requisite non-historical criteria for moral responsibility.
2. An evolved evaluative scheme is authentic if it results from acceptable modifications to a prior authentic evaluative scheme (which itself spells out what acceptable modifications are).

These two lead to their general account:

NH: If an agent's evaluative scheme at time *t* is either his initial authentic evaluative scheme or is an evolved scheme resulting from acceptable modifications to some prior authentic evaluative scheme, then his evaluative scheme is authentic.

As long as there is an approximate correspondence between what Haji and Cuypers have in mind with what constitutes an evaluative scheme, and what Mele has in mind with what constitutes critical control capacities, the two accounts are similar with respect to how an autonomous agent must change in a way that preserves autonomy: any changes to his psychological basis for evaluative reasoning must occur in light of that psychological basis for evaluative reasoning. However, while Mele's account identifies innate and early childhood acquired preferences, in so far as they are unshedtable, as inauthentic, Haji and Cuypers contend that such elements are by default authentic. Both accounts

⁸⁴ Haji (2013), 194.

deem the control capacities/evaluative schemes of instant agents as by default authentic, Haji and Cuypers because their unsheddable values are relationally authentic and part of their initial schemes, and on Mele's terms this results from all of an instant agent's various unsheddable values being instilled at the moment they get their control capacities—if such values had been implanted *prior* to having such control capacities, they would be inauthentic.

It is useful at this point to contrast Haji and Cuypers's account with how Mele deals with a case of religious indoctrination:⁸⁵

A religious fanatic initially conducts his child's religious instruction in the same matter-of-fact tone in which he teaches the child about mundane matters: there are birds, bees, bicycles, and buildings, and there is God, a devil, heaven, and hell. Then he teaches the child that there are a few extraordinarily evil people who "reject" God—the dreaded atheists—and that they will burn in hell forever, a prospect he graphically portrays. For good measure, he "informs" the child that people who even entertain doubts about his religious teachings also burn eternally. Suppose that the man does a thoroughly effective job, and the child—owing primarily to a deep-seated fear of eternal damnation—grows up with firmly held religious convictions and values that, even as an adult, he is practically unable to shed. Suppose further that very young children have no capacity for controlling what religious doctrines and values they accept, no capacity for making up their own minds about such matters. Then, it might be said, the father has compelled the child to have certain pro-attitudes without bypassing the child's capacities for control over his mental life: one cannot bypass nonexistent capacities.

Peculiarly, Mele sees the principle that "one cannot bypass nonexistent capacities" as implausible, and goes on to defend why the child's capacities *have* been bypassed, whereas this is essentially what Haji and Cuypers endorse to preserve the authenticity of innate and early acquired unsheddable values. This is most likely because Mele is inclined to see the child's religious views as inauthentic, and thinks he must identify some historical feature that explains this, writing:⁸⁶

[E]ven if the very young are not capable of making up their minds then about theological matters, they might then have the capacity to develop into individuals who would make up their minds about such things. For example, prior to his indoctrination, the child at issue might have had the capacity to become a person who would shape his own attitudes about theology later in life. If so, that capacity was bypassed— and, indeed, destroyed.

I believe this solution has severe complications. First note that the situation is underspecified. The

⁸⁵ Mele (1995), 167.

⁸⁶ Ibid., 168.

child's religious views must be, using Haji and Cuyper's terminology, either relationally inauthentic or not; that is, they either subvert responsibility for non-historical reasons (perhaps they are like an irresistible desire, or they mark the impairment of some sort of open-minded capacity to consider alternative religious viewpoints that is necessary for autonomy) or they do not. If they do, then there is a clear non-historical feature that explains why the child's views are inauthentic (no historical explanation is needed). I think this is the most likely scenario, where Mele believes (and I agree) that the indoctrination undermines responsibility because it actually impairs the child from ever deploying any critical scrutiny or rational rejection of the indoctrination, but if that is true then Haji and Cuyper's account can perfectly handle it. But if they don't, and the child's religious views are in the same category as Cuyper's O-unsheddable affinity toward classical music, then Mele is essentially saying that the religious indoctrination undermines responsibility because it prevents the child from becoming the person he would have without the religious intervention. But *any* significant circumstantial influence changes the sort of person a child will otherwise become, and so this would amount to the implausible view that every circumstance, save the complete absence of environmental factors, undermines autonomy. Thus, it seems the principled way to handle Mele's case is to say that either the religious indoctrination is so strong as to subvert autonomy for non-historical reasons, or it does not impair autonomy (since if it did, so would the sort of innate and early acquired preferences that seem essential to preserve as authentic in any theory of responsibility that is not utterly revisionary).

4.4 The Problem of Arbitrary Differences

At its core, Haji and Cuyper's account is Mele's account with the addendum that unsheddable values acquired before one has the requisite control capacities are default authentic, rather than necessarily bypassing and so inauthentic. This has the advantage of being able to account for why our innate or early acquired preferences do not actually undermine our freedom.

However, this account yields results that I believe are implausible. It appears that seemingly

arbitrary differences in when manipulation implants unsheddable values can result in significantly different responsibility ascriptions. If this is true, the objection is general enough to threaten the prospects of any negative historicist account. To see why, reconsider a condensed version of McKenna's Suzie Instant:

Suzie Instant: A god goes to some world and creates an instant agent, Suzie Instant, who satisfies [pick-your-favorite-non-historicist requirements for moral responsibility] and whose evaluative scheme contains an unsheddable value that calls for A-ing. Suzie Instant is immediately given a choice between A-ing and B-ing, where to B would be to violate her unsheddable value. Suzie Instant A-s.

Not only do Mele's and Haji and Cuyper's negative accounts all deem Suzie Instant free, it seems that *any* negative historicist account must do so. After all, the structure of a negative historicist account will include the specific histories that an agent must lack to autonomously possess his values. We are free to stipulate that instant agents like Suzie satisfy any non-historicist aspects of the account, and as an instant agent, she by definition lacks any past that could threaten responsibility.

It seems relatively noncontroversial that whatever evaluation we make of Suzie Instant should be symmetrically applied to:

Impulsive God-1: The god goes to another planet with the intention of creating an instant agent that satisfies the requisite non-historical criteria, and who has an unsheddable value that calls for B-ing. However, a nanosecond before he creates the instant agent he changes his mind, and creates an instant agent identical to Suzie Instant, Suzie Instant* who has an unsheddable values that call for A-ing. She A-s.

There is, as noted previously, a longstanding tradition within the compatibilist literature of denying the relevance of a manipulator's intentions with respect to whether some intervention undermines autonomy. This is for good reason, since if manipulation subverts autonomy it should be in virtue of the *effects* it has on one's psychological basis for practical reasoning, and intentionality apart from its downstream effects is not the sort of thing that could intrinsically have such effects. But we need not even appeal to as noncontroversial a principle as this in explaining why we should treat *Instant Suzie* and *Impulsive God-1* symmetrically; we should because *Impulsive God-1* could be the backstory of *Instant Suzie*. We may add whatever additional details we wish to *Instant Suzie* while preserving the autonomy

of the instant agent being created, so long as those additional details don't contradict the original conditions.

But contrast the instant agent in *Impulsive God-1* with the one in:

Impulsive God-2: The god goes to another planet with the intention of creating an instant agent that satisfies the requisite non-historical criteria, and who has an unsheddable value that calls for B-ing. However, a nanosecond *after* he creates the instant agent he changes his mind, and changes the instant agent to one identical to Suzie Instant, Suzie Instant**. She has an unsheddable value that call for A-ing. She A-s.

Is there a basis for making asymmetric judgments about the moral responsibility of the instant agent in *Impulsive God-1* and *Impulsive God-2*? Any such distinction would seem arbitrary. The only difference between the two is that in the former the god decides to make Suzie Instant* a nanosecond before creating her, and in the latter the god decides to change Suzie Instant** a nanosecond after creating her. Likewise, the only difference between *Impulsive God-2* and the original *Instant Suzie* is that in *Impulsive God-2* Suzie Instant** has, for only a moment, just one nanosecond, the unsheddable value to B before the god changes it to an unsheddable value to A. If Suzie Instant owns her values in a way that she can authentically act from them, it would seem arbitrary to think that such a small difference in time of having different values could result in Suzie Instant** not also being free.⁸⁷

To bolster this judgment, we can contrast:

Sleeping Suzie-1: The god goes to another planet with the intention of creating an instant agent who is asleep, and who has the unsheddable value to B. But right before he creates her, he changes his mind and creates Sleeping Suzie, who is identical to Suzie Instant except she is created asleep, and will only face the decision to A or B once she awakens. Eight hours later, she awakens and A-s.

Sleeping Suzie-2: The god goes to another planet and creates Sleeping Suzie*, identical to Suzie Instant** (from *Impulsive God-2*) when she was originally created (so Sleeping Suzie* is made with an unsheddable value that calls for B-ing) except she is created asleep, and will only face the decision to A or B once she awakens. At some point during the night, the god substitutes her unsheddable value to B with an unsheddable value to A. Sleeping Suzie* awakens and A's.

I take it to be noncontroversial that if Instant Suzie is free, so is Sleeping Suzie; whether an instant agent is created asleep or awake makes no difference to autonomy. Yet I find it extremely bizarre to

⁸⁷ Levy (2011), 103.

think that Sleeping Suzie is free but Sleeping Suzie* is not. While it is true that the difference is that Sleeping Suzie*'s evaluative scheme is *actual* and *existing* before it is changed while Sleeping Suzie's is merely potential (as the god changes his mind before he creates her), it is far from obvious to me that whether an evaluative scheme already exists in this way is relevant to autonomy. Now, if we ask *why* sleep and arbitrarily small time scales, specified in the contexts above, do not make a difference to autonomy, it seems that this is because our intuitions about moral responsibility do not track whether robust capacities for control simply *exist*, but whether they are also *functioning*—and with respect to *exercising* one's capacities to critically assess values, sleeping and extremely small time scales are analogous to non-existence. This idea is the subject of the following chapter.

But perhaps the intuition that Suzie Instant** should be judged symmetrically with Suzie Instant* (and in so far as she is relevantly similar, also Sleeping Suzie*) might be contested.⁸⁸ After all, the distinction between creating and affecting persons is widely seen as morally significant in other contexts, and thus the burden would naturally fall on the person arguing its irrelevance here. Haji and Cuypers, considering a radical reversal following instantaneous agency, remark:⁸⁹

To elaborate, suppose that moments after her 'birth', Suzie* falls victim to Ann/Beth style induction. We see little reason why [non-historicists] should deny that the manner in which Suzie* acquires the 'new' components of her evaluative scheme has distinct implications for responsibility even though pre- and post-manipulated Suzie* are equipped with the psychology that both [non-historicists] and externalists would deem sufficient for responsibility but for its odd provenance. We have registered that [non-historicists] acknowledge that factors such as coercion that may affect freedom, or factors such as deception that may compromise epistemic requirements of responsibility, may well be responsibility-subversive. But then there is no reason for [non-historicists] to resist the view that factors that compromise agency requirements of responsibility—factors, for instance, that threaten normative agency—may also be responsibility-subversive. In the initial scenario involving her 'birth', there is no question about subversion of agency; actions that causally arise from Suzie*'s authentic initial evaluative scheme satisfy [proper engagement with the action]. In the latter scenario involving global induction, there is an obvious concern with the subversion of agency; actions that causally arise from Suzie*'s modified evaluative scheme fail to respect [proper engagement with the action].

It is difficult to tell whether Haji and Cuypers would report a different intuition about Instant Suzie**.

⁸⁸ Neil Levy (2011, 103) considers a similar instant agent who is immediately changed, and shares my intuition.

⁸⁹ Haji and Cuypers (2007), 361-362.

They make this point in the context of replying to the non-historicist argument that if instant agents are indeed free, this is evidence for a non-historical view. Their thought experiment is meant to illuminate the point that while it is true that for agents with no past, their history has no bearing on their responsibility, this fact does not suggest that for agents *with* a past, their history has no bearing on their responsibility—and this is a point with which I agree. But while the phrase “moments after birth” does seem suggestive of a small time scale, I cannot tell if Haji and Cuyper are imagining Suzie*, for a few moments, acting from her initial evaluative scheme, and then being globally inducted, or if she is globally inducted before she even has a chance to begin acting. If it is the latter, then they do have different intuitions in considering differences that I think are irrelevant to autonomy, but I will have to wait until the following chapter to address this. But if it is the former, then this would be compatible with the observation that it matters not whether an evaluative scheme merely exists, but whether it is also operative or functioning.

I will expand on this principle far more in the next chapter, but for now it suffices to say that in so far as Sleeping Suzie* and Instant Suzie** should be judged symmetrically with Instant Suzie, despite the manipulation that occurs after they pop into existence, there is a big problem with Mele’s and Haji and Cuyper’s accounts. In Mele’s language, both instant agents are created with capacities for control that are bypassed since the change of an unsheddable value to B with an unsheddable value to A does not in any sense go through the agents’ evaluative frameworks. In Haji and Cuyper’s language, both agents are created with initial evaluative schemes of the sort that allows them to act freely. But in both cases, these evaluative schemes are modified in inauthentic ways, and thus the actions issuing from these modified elements cannot be free. Now, the inability to account for test cases is not a fatal complication with any account, and the negative historicist is welcome to attempt to explain away the intuition that both are just as autonomous as non-manipulated instant agents.

However, an alternative to explaining away counterintuitive implications of negative

historicism is to first see whether some other account can accommodate these test cases. While positive historicism straightforwardly can (none of the instant agents have satisfied the ownership condition we previously discussed), we have reviewed its own complications. Is there some alternative to the negative and positive historical accounts we have encountered? Interestingly, the objections I have leveled against negative historicism seem to suggest that positive and negative historicism encounter complementary complications on two sides of the same coin: both accounts deem unfree agents who intuitively are (instant agents in the case of positive historicism, and agents like Sleeping Suzie* and Suzie Instant** in the case of negative historicism). But they do so by falling on opposite sides of how demanding a conception of autonomy they have. Positive historicism seems to set the bar too high for what it means to authentically own one's values, resulting in an overly restrictive and so overly exclusive conception of autonomy that instant agents like Suzie just can't meet. On the other hand, negative historicism seems to set the bar too low, resulting in an overly liberal and so overly inclusive conception of autonomy, deeming intuitively innocuous manipulations of agents like Sleeping Suzie* and Suzie Instant** as responsibility-subverting. In other words, positive historicism makes it too hard to take ownership, and negative historicism makes it too easy. We may be at a dialectical stalemate, where we need to choose between two incompatible but desirable conditions. On the other hand, there may be a middle ground that can capture the upshots of positive and negative historicism without their counterintuitive implications. It is this middle ground on the ownership condition that we will now explore.

CHAPTER V: Operant Historicism

5.1 Operativity

Abstractly, one's evaluative scheme—his psychological basis for acting—is composed of myriad evaluative elements (beliefs, desires, deliberative principles, etc...) that holistically determine how he will execute practical reasoning. On Haji and Cuyper's account, there is only one dimension by which we judge whether an initial scheme is authentic, and that is whether it is sufficiently complex. Let us call this measure of structural complexity and soundness *robustness*, so that an evaluative scheme is authentic in so far as it is composed of whatever constellation of evaluative elements, at the requisite level of robustness, necessary to meet the non-historical criteria of moral responsibility. The historical aspect of their account, in turn, comes from how a person's different evaluative schemes over time relate to one another.

Our discussion of instant agents who are, in their very first moments of existence, manipulated into becoming completely different people, seemed to beckon for another dimension that is relevant to moral responsibility: not just how robust each evaluative element is, but another criterion which I previously hinted at but did not investigate: *operativity*. That is, I am suggesting that the authenticity of some evaluative element is *pluralistic*, and depends not only on its robustness, but also whether it has been used or exercised, or how *operative* each evaluative element is. Our values inevitably play a causal role in our behavior, and operativity tracks the degree to which they have been causally implicated in our behavior; when we, over time, engage in practical activity, which is meant to denote the broadest possible means through which we express our values, those values become more operative. Certainly, robustness and operativity usually go hand in hand. As we express our values more and more, they become more and more complex. However, the two can come apart. Suzie Instant can have the most robust evaluative elements we wish to stipulate, but they are completely inoperative the moment she is created (then, as she begins reflecting on her values, or acting from them, or some other form of practical activity, her values become evermore operative). By contrast, a child or young teen who has

gone for a year making free choices will have evaluative elements more operative than Suzie Instant's, but such elements may not be as robust (though, by stipulation, they are robust enough to meet the non-historical criteria for moral responsibility).

This suggests a supplementary dimension to our account of moral responsibility—one which depends not only on what elements are in an evaluative scheme and what modifications led to that current evaluative scheme, but also on the degree to which elements in the evaluative scheme have been exercised through practical activity. To draw a parallel to Haji and Cuyper's developmental description of acquiring moral responsibility, the equivalent in this revision of the account would be something along the lines of the following: the first stage of moral agency is identical to Haji and Cuyper's, where a newborn infant up through early childhood lacks the cognitive complexity to have what might be called an initial scheme, and at best has proto-values that will later figure into that scheme. At this stage, the child is not morally responsible for his behavior. At some point, the child's (relationally authentic, to borrow from Haji and Cuyper's terminology) proto-values become robust enough and integrated with one another to form an authentic initial evaluative scheme. However, at this point, although certain elements are robust enough to form the initial evaluative scheme, it is inoperative, because the child has not yet expressed any of the values embedded in the scheme through practical activity. Over time, the values in the inoperative evaluative scheme are causally implicated in the person's thoughts and actions, becoming more and more operative. And at some point, an evaluative element becomes sufficiently operative that when the person expresses that value, he is morally responsible for it. Given that sufficient practical activity is required for a person to "take ownership" over his value so as to be morally responsible when acting from it, while manipulation that strips or replaces operative values is freedom-undermining, manipulation that strips or replaces *inoperative* values only removes values that a person hasn't taken ownership over, and thus is not

freedom-undermining.⁹⁰

We can appreciate the conceptual upshots this has with respect to the complications with negative historicism. Recall that negative historicism returned responsibility-subverting verdicts for people like Sleeping Suzie* and Suzie Instant**, who have their evaluative schemes changed arbitrarily soon after they pop into existence. According to this account, what I will call *operant historicism*, such interventions are not bypassing, and so freedom-undermining, since elements of an evaluative scheme can only be bypassed if they are (sufficiently) operative, and the evaluative elements in such instant agents are robust but insufficiently operative:

Operant Historicism (general structure): An agent performs a directly free act and is directly morally responsible for it only if any O-unsheddable⁹¹ values playing a role in the production of her action are sufficiently operative, so none bypassed her capacities to critically assess and engage them from the abilities she possessed.

At first glance, it is not clear how “sufficient operativity” squares with the historical requirements in positive and negative historicism. I clarify this in 5.3, but in order to do so, we first need to examine one of operant historicism’s advantages: its ability to account for responsibility after manipulation.

5.2 What Mele Did Consider: Can Manipulated Agents Regain Their Freedom?

In articulating sufficient conditions for non-responsibility, Mele makes an interesting remark:⁹²

NF. An agent who performs an overt action A does not freely A, if (1) he expresses unsheddable values in A-ing; (2) owing directly to those values, he could not have done otherwise than A in the circumstances (on a compatibilist reading of “could have done otherwise”); (3) those values were very recently produced in a way that bypassed his capacities for control over his mental life by value engineering to which he did not consent and are seriously at odds with autonomously acquired values of his that were erased in the process; (4) he retains no preexisting value that is promoted by his having the unsheddable values he expresses in A-ing; and (5) A is the first overt action he performs on the

⁹⁰ The reader may object to my terminology and assert that one could hardly be said to have a value if he has never acted from it. This is a semantic point that does not bear on the substance of my argument. If it is true that we cannot be said to have a value or desire we have never acted on, then we can rephrase the criticism of negative historicism and say that it allows us to take ownership over evaluative elements that are not values or desires. Thanks to Neil Levy for raising this potential concern.

⁹¹ Again, I am assuming that the main difficulty posed for accounting for the authenticity of especially innate and early childhood unsheddable values is accounting for the O-unsheddable ones, which are harder to account for than O-sheddable ones. If one believes that even O-sheddable values can be freedom-undermining, then whatever is used to soundly account for the authenticity of O-unsheddable values can presumably account for the authenticity of O-sheddable ones too (perhaps with even fewer requirements, since they are less buried in an agent’s evaluative scheme).

⁹² Mele (2006), 170.

basis of his new values.

Mele's qualification in (3) that the bypassing be "very recent" and his qualification in (5) that the action is the "first overt" action suggest that he does not see such manipulation as clearly *permanently* subverting responsibility—maybe the agent will always be unfree, but maybe he won't. Unfortunately, Mele never elaborates on this point or provides a principle that would explain this. Moreover, while (5) stipulates the possibility of what we may call *re-emergent autonomy*, it is ostensibly not possible according to his account. Recall that Mele's sufficient conditions for being compelled to possess a pro-attitude are essentially that the pro-attitude is practically unsheddable and bypasses S's modest capacities for control. So long as the pro-attitude remains practically unsheddable, it will remain inauthentic, and no amount of time that doesn't simply erode the unsheddability of the value will allow for it to become authentic. Nor can Haji and Cuypers' account make sense of this type of re-emergent autonomy, since an evolved scheme can only be authentic based on acceptable modifications from a previous authentic scheme. According to the letter of the account, once an evaluative scheme is bypassed and becomes inauthentic, further modifications to that scheme are modifications to an inauthentic scheme, and so again such evolved schemes will be inauthentic. Since the kind of manipulation we are discussing here renders an evaluative scheme inauthentic, their account, like Mele's, will be deficient.⁹³

This seems problematic, because while manipulation, stipulated properly, strips autonomy in the immediate moments after it occurs, it seems at least *prima facie* plausible that if enough time goes by, the manipulated agent may eventually take ownership over his initially inauthentic values. This

⁹³ McKenna's positive historicism does better with respect to re-emergent autonomy, where a manipulated agent still can reclaim ownership over manipulated values by having the right opportunity for rational reflection. His positive requirement emphasizes that the opportunity to assess must operate on "abilities that she possessed," which is presumably to avoid the objection given to identification style nonhistoricist accounts that an agent can be manipulated to identify with his actions. If the opportunity to critically assess values does not depend on abilities one already has prior to manipulation, but could itself come from having sufficient opportunity to assess from any critical capacities, regardless of their etiology, then the agent just needs to "start over" after global manipulation, positive historicism can account for re-emergent autonomy. Though the way it does so would still be problematic, as I argue next.

intuition seems to not depend entirely on an agent gaining the ability to shed such values over the longer period of time. After all, consider:

Manipulated Axe. Suppose that humans normally become free and morally responsible agents at around age eight. Axe is an ordinary eight year old who developed an authentic (nonhistorically freedom-preserving) initial evaluative scheme at eight, and has been freely acting for a few months. He is subjected to global manipulation that completely changes his evaluative scheme, replacing many of his O-unsheddable values with different ones that will remain O-unsheddable for the rest of his life.

Suppose Axe lives an ordinary human life of about ninety years. While it seems plausible that he is not responsible for some period of time after global manipulation, it would seem strange if, if people are indeed morally responsible for their actions, manipulation that happened in his early stages of autonomy blocked him from being morally responsible with respect to the manipulated values for the rest of his life.⁹⁴ It seems unprincipled that Axe's initial scheme—which could have set him on the path to becoming a murderer, or politician, or plumber, or *any path* the evaluative scheme of which meets the requisite non-historical criteria—was default authentic, but as soon as that scheme is modified in an unacceptable way, even if the path it specifies is quite close to his original one relative to the vast range of initial possibilities, his responsibility is subverted for the rest of his life. This would even be the case if humans develop massive life-lengthening technologies, and Axe lives for another 100,000 years. This seems to place too great weight on one's initial character endowment and circumstances, and does not appear to be a plausible account of moral responsibility.

Operant historicism, on the other hand, does have the conceptual resources to account for re-emergent autonomy. Suppose that there is a scale for measuring how operative an evaluative element is, the value of which increases as that element is causally implicated in practical activity. And imagine, for the sake of argument, that an evaluative element must have an operative value of at least 100 in order for one to freely act from it. If that evaluative element is erased and replaced with one seriously

⁹⁴ I am assuming that an otherwise free agent who has only one evaluative element finagled only loses autonomy when acting from that finagled element (given that that element does not influence the effects of his other values). Of course, our values are much better understood as forming a web of myriad interrelations, though I ignore this complication here.

at odds with it, then the person does *not* act freely from this value (the intuitive result), but the explanation is that it is now a 0 with respect to operativity. If, through practical activity, the person continually acts from that value, eventually, according to one version of operant historicism, when it reaches 100 in operativity again, the autonomy re-emerges.

5.3 The Road to Responsibility: Why This Isn't Just Positive or Negative Historicism

Before I raise an illuminating objection to operant historicism, a minor point needs to be made. Operant historicism, in so far as it requires evaluative elements to be operative, does have a positive historical constraint. Evaluative elements by stipulation become operative through an agent, over the course of his history, engaging in practical activity. But this is far different from what McKenna imagines in even his most general form of positive historicism, and he in fact *excludes* the possibility of invoking operativity for reasons I will now criticize. McKenna understands positive historicism in non-practical terms—that is, terms that do not require action. On this understanding, an agent just needs to have a sufficiently complex capacity for rational reflection which he need not even exercise; he only needs the *opportunity* to reflect on his values in order to take ownership over his actions. By contrast, the positive aspect of operant historicism *does* require practical activity to make inoperative elements of an evaluative scheme sufficiently operative. And this is a requirement that McKenna directly intends to preclude from his own account:⁹⁵

The phrase “the opportunity to critically assess, endorse, and sustain from abilities that she possessed” in PH must be interpreted so that the evaluative processes of assessment, endorsement, and sustaining involved *do not require action, and hence do not require free action*.⁹⁶ All they require is a proper degree of cognitive control or activity. So, all PH demands is that an agent have had the opportunity to assess her values in this more inclusive way.

McKenna qualifies his account this way in response to mentioning the regress objection to versions of positive historicism whereby some positive historical requirement is itself analyzable in volitional

⁹⁵ McKenna (2016), 99.

⁹⁶ Emphasis mine.

terms, making it seemingly impossible for someone to develop moral responsibility because every free act would itself require some prior free act. Thus, it appears that his reason for not requiring action is that by not requiring action, his account necessarily does not require free action, and thus will be immune from any regress concerns. However, this is a premature abandonment of action, or at least one done for unnecessary reasons. Action of course does not entail free action; and so as long as one is careful to not require a positive historicist conception of action that is analyzable in volitional terms, the regress objection can be avoided without throwing out practical activity altogether. And this is exactly what operant historicism does. The practical activity requirement does not say that a person must freely exercise his evaluative elements to take ownership over them, but only that he exercise them.⁹⁷ Since this is a precondition for moral responsibility, such practical activity need not be free, but still captures the intuition that one is not tethered from the start to his endowment, but must exercise his endowment to be free. In any case, McKenna's policy of avoiding a positive requirement cashed out in actional terms makes it clear that even though operant historicism has a positive aspect, it is unlike anything McKenna has in mind.⁹⁸

This having been said, we can clarify how this fits into the seemingly collectively exhaustive distinction of historicist positions by addressing an objection. The objection is that operant historicism is not a novel form of historicism, but is rather a vague account that covers the *disjunction* of positive and negative historicism, and when it is further specified, collapses into one of the two. We might think this for the following reason: I have left open *how* operative a value must be to meet the necessary operant historicist requirement. Either the value must be negligibly operative or not. If the value must

⁹⁷ Of course, behavior is just one type of practical activity. It is possible for someone, through an omission, to increase the operativity of some evaluative element relevant to the omission.

⁹⁸ Indeed, since rational reflection itself might be a form of practical activity, there is a sense in which McKenna's generalization of positive historicism is just an (implausible) subclass of operant historicisms—those that assert the only way to increase operativity is through the practical activity exhibited during rational reflection.

be negligibly operative or not even operative⁹⁹, then the account is just negative historicism by a different name. On the other hand, if the value must be more than negligibly operative, then the person must *have undergone practical activity*, in which case this is a positive historical requirement that an agent *have* a past that contained sufficient practical activity so as to make her O-unsheddable values sufficiently operative.

The problem with this objection is that it assumes that the degree of operativity of an evaluative element, in order for an agent to act freely from it for the first time, must be the same regardless of whether manipulation has taken place—but this need not be the case. I will discuss this more fully in the next section, but for now it is important to outline when discussing operant historicism. Suppose that ϕ is the operativity threshold: the value an evaluative element must first reach for an agent to act freely from it, and that for some evaluative element E, $\phi=100$. Suppose the agent, who has freely acted by E for some period of time, is manipulated and E is replaced with a significantly opposing value E', so E' has an operativity value of 0. The objection assumes that E' is authentic once it reaches level 100 again, but operant historicism need not be so rigid. After all, at the point E is replaced, it might have an operativity value of 200, and an operant historicist account might require not that the manipulated value return to a value of 100, but rather 200. The motivation for this stems from contrasting the original case of manipulated Beth with an instant agent who is created to be, like manipulated Beth, Ann's value twin. Assuming $\phi > 0$, Ann is morally responsible for her immediate actions following the time that Beth is manipulated and the instant agent is created since her values are operative and have been practically exercised for her many years of ordinary life. Both manipulated Beth and the instant agent have evaluative schemes that are as robust as Ann's, but which

⁹⁹ It may at first appear bizarre to invoke the notion of operativity only to then specify that no operativity is needed for an action to be free. After all, isn't the point of invoking operativity to highlight the importance of practical activity for moral responsibility? It is in addressing the objection that operant historicism is either negative or positive historicism that we can appreciate why operativity is still important even if one's *initial* scheme need not be operative to ground responsibility.

are completely inoperative, and they begin the road to responsibility by exercising their evaluative elements until they are sufficiently operative such that they are apt candidates for blame and praise, and let us assume that they encounter identical life experiences. The contention is that it is not immediately obvious that manipulated Beth and the instant agent *become responsible at the same time*. There is nothing *taken away* from the instant agent whereas an evaluative scheme that Beth took ownership over for many years of normal life was completely reversed. It seems *prima facie* plausible that Beth will not become responsible until later than the instant agent for this reason, and so allowing the level of sufficient operativity to be a function of *when manipulation occurs* in an agent's history accommodates the intuition that the *more an agent has acted on his values, the more his autonomy has been undermined when they are replaced* (and so the more it takes for him to take ownership over his new values). Indeed, this is very clearly McKenna's intuition about the difference in the Beth Passive/Active case, but operant historicism actually gives a principled explanation of it.

On the account just suggested, there is a linear relationship between how operative an evaluative element is and how operative a finagled element needs to be in order for it to be authentic. Of course, the operant historicist has conceptual resources far more flexible than this. He could, for instance, propose a more exponential relationship, where the more operative an evaluative element is, the operative value a finagled element must reach to restore authenticity increases increasingly. He could also, for instance, assert some threshold value, where an evaluative element has become so operative and so deep a part of a person's identity that if it is finagled and replaced he is permanently non-responsibility for actions issuing from it. I will not explore these various possibilities in this thesis. I raise them to highlight the fact that operant historicism, in its general form, includes an *operativity function* that takes as input the operative value of the original evaluative element for which the finagled value conflicts and outputs the operative value that the finagled element must reach in order to be authentic. Thus, if ϕ is the operativity threshold a particular evaluative element constituting an O-

unsheddable value must reach in order for it to be authentic; O_T is the operative value of that element at the time of action; $f(O_H)$ is the operativity function that is dependent on the operative value of the original element that the finagled value replaces or conflicts with; and μ^{100} is the proportion measuring the degree to which the original value and manipulated value conflict.¹⁰¹

Operant Historicism: An agent performs a directly free act and is directly morally responsible for it only if for any O-unsheddable value X of operativity O_T that replaced and opposes an evaluative element of operativity O_H playing a role in the production of her action, $O_T \geq \phi$ and $O_T > \mu \cdot f(O_H)$.

That there can be a divergence between when a value becomes free and when a finagled value becomes free is particularly useful with instant agents. In so far as the operant historicist believes that instant agents are responsible for their initial free actions, he will assert that $\phi=0$. And if he wishes to uphold this conclusion without a commitment to the freedom-undermining effects of manipulation that occurs nanoseconds after creation (the apparent flaw with negative historicism), he will explicate the operativity function such that a finagled value that replaces an original value with a low operative value needs minimal or even no operativity to be authentic.

This also lets us appreciate why this does not easily categorize as negative or positive historicism. Suppose the operant historicist asserts an account according to which $\phi=0$ (for reasons we will consider in the next section), and the operativity function is linear. In this case, Suzie Instant and other instant agents with sufficiently robust evaluative schemes would be free for their initial actions, since there is no requirement that actions issuing from their initial schemes be at all operative. However, if after her first week of life, Suzie Instant's O-unsheddable desire to, say, not steal is

¹⁰⁰ The μ proportion derives from the natural thought that the less opposed some implanted value is to an agent's already autonomously values and reasons, the less it undermines responsibility. If I have a desire to donate my money to breast cancer charities, and you replace it with a desire to donate to ovarian cancer charities, this seems less responsibility-subverting than removing the desire altogether. The μ proportion simply gives the operant historicist a way of accounting for this, though the proportion will not be further analyzed.

¹⁰¹ This is itself a probably too simple account. When manipulators implant an unsheddable value, it in all likelihood conflicts with more than just the evaluative element it replaces. Hence, a fully fleshed out account will probably have some integral over all of the evaluative elements (with their respective operative values and degrees of opposition) that the manipulation opposes.

replaced with a desire to steal, and she has already had many opportunities to steal that she's refrained from, she must undergo as much practical activity stemming from that newly implanted desire before she is responsible for her stealing. Because of this, operant historicism is compatible with instant agents, whose evaluative elements by stipulation are wholly inoperative, being free, and also for such instant agents, *after they have exercised those evaluative elements*, to be (initially) unfree when their new evaluative elements are inoperative. This means that such an operant historicist account would not require that one has a past if one has never been manipulated in a certain way, but if one has, it must be a past of a certain sort. This account *would be a negative historical account with respect to unmanipulated agents, and a positive historical account with respect to manipulated agents*. We can call this:¹⁰²

Negative Operant Historicism: An agent performs a directly free act and is directly morally responsible for it only if for any O-unsheddable values X of operativity O_T that replaced and opposes an evaluative element of operativity O_H playing a role in the production of her action, $O_T \geq 0$, and $O_T > \mu \cdot f(O_H)$.

By contrast, if the positive historical requirement applies even before an agent has been manipulated (some practical activity is required for any evaluative element to be free), we get:

Positive Operant Historicism: An agent performs a directly free act and is directly morally responsible for it only if for any O-unsheddable values X of operativity O_T that replaced and opposes an evaluative element of operativity O_H playing a role in the production of her action, $O_T \geq \phi$, $\phi > 0$, and $O_T > \mu \cdot f(O_H)$.

Since positive operant historicism has a positive historical requirement for everybody, it is a form of positive historicism. However, negative operant historicism does not have a positive historical requirement for everybody—namely, agents without a past, and agents who have never been manipulated—but rather a positive historical requirement for people with a certain sort of history (that involves manipulation). Indeed, negative operant historicism essentially says that whether moral responsibility has positive or negative constraints is *itself a historical question*.

¹⁰² As this hopefully makes clear, I am not setting up operant historicism as an account that is in direct competition with positive and negative historicism (i.e. an account that stands in contrast with a positive historicism that places independent weight on rational reflection). Rather, operativity is a feature that I think the best account of historicism will invoke (and hence specifying operativity as a necessary feature for moral responsibility in operant historicism, but not a sufficient one), leaving open whether we ultimately want to combine operativity with the elements of positive or negative historicism.

5.4 Comparing Operant Historicism

If one is attracted to historicism, there seem to be good reasons for being an operant historicist of some sort—varieties of operant historicisms capture the conceptual upshots of the other types of historicisms, and avoid their respective pitfalls. The closest analog to negative historicism is negative operant historicism, which, to stress, has a positive historical requirement in some agential histories. Like negative historicism, and in contradistinction to positive historicism, negative operant historicism allows us to deem instant agents like Instant Suzie as morally responsible for their first free acts. However, while negative historicism had no recourse to the objection that it generates different responsibility judgments across arbitrary differences, negative operant historicism does not, because one's critical capacities have not genuinely been bypassed if one has not actually used them.

On the other hand, the closest analog to McKenna's positive historicism is positive operant historicism, which also has a more robust "taking ownership" condition compared to negative historicism. However, the major problem with McKenna's positive historicism, I argued, was that an opportunity for rational reflection, given that it doesn't actually have a chance of shaping one's values, doesn't really seem any more plausible a constraint than some non-historical identification criterion. On the other hand, it seems much more plausible that the actual undertaking of practical activity can affect the extent to which one has taken ownership for his values in a way that the opportunity for rational reflection cannot. Operant historicism takes the common thought that you can best tell what a person values not by what he says but how he behaves, and positive operant historicism elevates the importance of this into a positive requirement.

Operant historicism also seems to account for *why* rational reflection is so important. In operant historicism, there is no presupposed central set of capacities or fundamental values in light of which new evaluative elements must relate in the right way. There are simply more and less operative elements in one's evaluative scheme. Naturally, however, there are certain elements we *use* far more

frequently in practical activity than others. My evaluative element that is my desire for ice cream is exercised rather infrequently in comparison to what evaluative elements spell out my general approach to practical deliberation and other deliberative principles. Such elements will become far more operative, and hence if they are manipulated and replaced, this will be more freedom undermining than changing my ice cream taste is, if at all. But the *reason* for this is not because such capacities are stipulated to be privileged, but because they happen to be the ones that are most core in practical activity. Thus, operant historicism offers an additional explanation for why there is something particularly dark about the finagling of one's core rational capacities.

There is another advantage to operant varieties of historicism. It is common accepted that moral responsibility is a *scalar* notion. If some human beings are fully responsible for their actions, at some point during their development they were less responsible, and even earlier they had only a minimal amount of responsibility. This poses a challenge for any compatibilist view that has binary conditions that can either be met or not, in which case we could presumably lack responsibility entirely for the first decade of our lives and then one day just assume full-blown responsibility. Operativity by definition is something that comes in degrees as evaluative elements are more and more exercised during practical activity. Thus, operant historicism offers a new conceptual tool for accounting for the scalar nature of moral responsibility.

There remains the question of which of negative and positive operant historicism is more plausible (the investigation of which will reveal further advantages of operant historicism). At first, the issue seems to boil down to our intuitive verdict on instant agents. If one finds that instant agents are responsible for their actions immediately following extemporaneous creation, he will set the operativity threshold (for non-manipulated agents) to zero and opt for negative operant historicism. If he holds the judgment that instant agents are not responsible, then he will set the operativity threshold higher to explain why instant agents are not responsible. Our discussion of operativity

pushes this intuition-clash forward in at least three ways. First, in so far as one is persuaded by the challenge of irrelevant differences against negative historicism, this seems to provide evidence in favor of the non-responsibility of instant agents; pointing to operativity is pointing to the importance of having a history whereby one exercises his values to take responsibility for them (the negative operant historicist will, of course, respond by saying that pointing to operativity only points to the importance of having a history once one has had his values finagled, but this seems to be a much less natural way of explaining the thought experiments).

Second, appreciating how evaluative elements become more operative gives us some reason for discounting intuitions that instant agents are responsible *despite having inoperative schemes*. This is because instant agents are often stipulated as being created with extremely *robust* evaluative schemes full of collections of false memories and falsely anchored identities. Because more complex evaluative elements enable a wider range and deeper scope of practical activity, the evaluative elements of such instant agents will become operative very quickly (we can imagine Suzie Instant is created mid-thought, considering to act on some value; since thinking about exercising one's values is a form of practical activity, she pops into existence already increasing the operativity values of many of her evaluative elements). By contrast, while I do have an intuition that instant agents like Suzie Instant are responsible, when I consider a non-historical duplicate instant agent of some child who is just committing his first free act, I have a much weaker intuition that such an instant agent is responsible—the positive operant historicist would explain the difference in such matters by appealing to the fact that the instant child has a much less robust evaluative scheme than Suzie Instant, and so will take a lot longer to increase the operativity value of his evaluative elements above the operativity threshold. Suzie's scheme is *so* robust that minimal practical activity makes her responsible; the child is permitted to “experiment,” so to speak, with his values before being responsible. And if the positive operant historicist takes this line and Suzie Instant's scheme is so robust that even *thinking* about expressing

some O-unsheddable value is sufficient practical activity to make it operative enough for responsibility, it looks like positive operant historicism can also account for the freedom of instant agents. I find settling subtle matters of what my intuitions are responding to to be extremely difficult, but it seems plausible that our intuition (if we have it in the first place) that instant agents are responsible depends on imagining such agents engaging in at least some practical activity (even if it is just reflecting on their falsely-informed values). I do not believe either of these points alone, or even both of them collectively, offers anything close to a decisive argument against negative operant historicism—I only mean to throw a few more considerations into the mix of this difficult debate.

Thirdly, the operativity function underscores an important consideration in how we should understand moral responsibility in the first place, which incidentally also bears on the relative merits of negative and positive operant historicism. As I mentioned earlier, it is plausible that the more operative an agent's evaluative elements are, the more his autonomy has been undermined when they get finagled, and consequently the more practical activity is required for re-emergent autonomy across finagled evaluative elements. If this is true, then there is a *floating threshold* for any evaluative element's reaching authenticity that is contingent on an agent's history, how operative that element has been previously, whether manipulation has taken place, and so on. If I first become responsible when acting from some evaluative element when it reaches an operativity value of 100, but I continue exercising it through practical activity until it reaches a value of 200 and is replaced by some engineered and seriously opposing value, that new value must exceed well beyond 100 in order to be authentic. However, this intuition is not universally shared, and we might be inclined to think there is a *fixed threshold* for any evaluative element's reaching authenticity; even when my value is replaced, I may need to start again from zero, but once it reaches 100 again I regain autonomy when acting from my finagled value. In the same way, if a person's IQ is 150 and increases over time but an organization only requires an IQ of 130 to join, should the person undergo a brain injury that temporarily reduces his IQ to 80,

he can rejoin the organization as soon as it passes 130 again, not only when it reaches the highest it was previously. An operativity function only makes sense with the former model; if the latter is true, there is no operativity function, just the operativity threshold.

The first thing to note about these two ways of understanding operant moral responsibility is that if one is attracted to a fixed threshold, this will present yet another strike against instant agents. An operant historicism without an operativity function has only the operativity threshold, and so if one wants to invoke operativity at all the threshold must be greater than zero. Thus, if we buy into a fixed threshold understanding of moral responsibility, we must be positive operant historicists; by contrast, a floating threshold view is compatible with both negative and positive operant historicism. Even if one does not share McKenna's intuition that manipulated Beth's responsibility is more undermined and will take longer to attain re-emergent autonomy than instant agent Ann or Beth Passive, I believe there are other reasons for favoring the flexible threshold view. Such reasons themselves emerge when we look deeper into Mele's concept of radical reversal.

5.5 What Mele Didn't Consider: How Radical Reversal Affects Radical Restoration

Consider Mele's variation of one of his protagonists, Chuck:¹⁰³

ONE GOOD DAY: For many years, Chuck has taken great pleasure in killing people. 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...In hardening his heart as he did, Chuck ensured that he had no values at all that could motivate a charitable deed...All this changed when manipulators got their hands on him. Over night, without Chuck's consent, they erased his bad values and replaced them with good ones that match Sally's. Shortly after he awoke, he started working with a local Habitat for Humanity crew in his neighborhood. When the work day ended, he drove around town for an hour and bought several boxes of Girl Scout cookies from every Girl Scout he saw—about fifty boxes in all. Then he delivered the cookies to a local homeless shelter...When Chuck fell asleep that night, the brainwashing was undone and Chuck returned to normal (for him); the manipulators were conducting a one-day experiment.

¹⁰³ Mele (2013), 159.

Mele intends this thought experiment to evoke the historicist intuition that Chuck is not responsible for his good deeds, and he adds the feature that Chuck is manipulated *back* to his original state to bolster this intuition. I agree that Chuck's reversion to normalcy illuminates that he is not responsible for his good deeds, but it also raises an important question: if agents' moral responsibility is undermined when their evaluative schemes are changed in certain ways, what happens with respect to moral responsibility when those manipulative changes are undone?

On the one hand, if an evaluative scheme is sufficiently at odds with another such that replacing it with the other is freedom-undermining, it would seem that replacing that new one with the original would be equally freedom-undermining. On the other hand, it also seems fair to say that in the second instance of manipulation, when the original manipulative changes are *undone*, one's rational capacities are not bypassed a second time, but in fact *restored*. The latter interpretation appears to be the one Mele favors, and indeed seems correct. After all, consider:

One Good Nanosecond: Chuck lives the same life pre-manipulation as in One Good Day, but instead changing his values toward more charitable ends for a day, the manipulators only change Chuck for a nanosecond and then, perhaps in a pang of bad conscience, change his modified evaluative scheme back to what it was originally.

It seems implausible that it could make any relevant difference to Chuck's responsibility if, after freely acting from his murderous evaluative scheme for many years, that scheme is temporarily—for a shorter time period than he can even form a thought—changed and that change is immediately undone. But if this is true, any differences between One Good Nanosecond and, for instance, One Good Minute (exactly the same as One Good Nanosecond except for the duration of the change) or One Good Hour, should be differences in degree rather than differences in kind. On Haji and Cuyper's model, it is hard to see how this restorative autonomy could be accounted for. That model already struggled with re-emergent autonomy, for the reason that a radical reversal creates an inauthentic evolved scheme, and further modifications to inauthentic schemes just result in other inauthentic evolved schemes. While McKenna's model could account for re-emergent autonomy so

long as the radically reversed person is given the opportunity for rational reflection, in these restorative cases the letter of McKenna's positive historicism does seem to allow for restorative autonomy, but only does so by treating it as just another case of re-emergent autonomy. Chuck, to act freely, now needs the opportunity for rational reflection. The problem, then, is that positive historicism fails to treat it as a special case of re-emergent autonomy where autonomy re-emerges as soon as the change is undone.

The obvious reply from the McKenna-style positive historicist is to revise his account so that any unsheddable values must have persisted through a history in which the agent had the opportunity for rational reflection, *unless he has already had that opportunity*. Since Chuck has already been able to evaluative his murderous proclivities from the *same abilities* (that now have been restored), the positive historicist can say that he does not need to start over when he returns to his pre-manipulation state. And perhaps the negative historicist can make a similar move, where evolved schemes are inauthentic if they result from modifications to inauthentic schemes, unless those modifications result in a scheme that is identical to a *previously authentic scheme*. While this addendum captures the intuition of restorative autonomy, the way it does so makes its account inherit a severe problem: what if Chuck is not restored to his former self immediately before the manipulation, but some time further back? Suppose that Chuck has been a morally responsible agent since age 10 and that Mele's manipulation occurs when he is 30. On the revised account just suggested, Chuck's autonomy is immediately restored not just if the manipulation is undone, but if his evaluative scheme is changed to that of his 20-year-old self—when, since he was a responsible agent at that time, he did have the opportunity for rational reflection of his unsheddable values. Indeed, the neuroscientists don't even need to turn the dial back too far in undoing the manipulation done at 30 for this sort of atavistic autonomy. On the revision just sketched, any manipulation that reverts an agent to some prior time in his life when (for positive historicists) he had the opportunity for rational reflection or (for negative historicists) he had an authentic scheme

will sustain autonomy, but such radical changes are clearly, if anything is, at least responsibility-undermining in the short term. Manipulators cannot just make me a value-twin to my former self, which could be so radically different from who I am now, without subverting responsibility in some cases.

To avoid this implication, the addendum to account for restorative autonomy must require that any unsheddable values persist through a history in which an agent had the opportunity for rational reflection, *unless his most recent autonomous self has had that opportunity*. On this proposal, after Chuck is manipulated at 30, only changes that rewind his evaluative scheme to moments just prior to the manipulation restore autonomy, and thus the positive historicist is not committed to some spooky atavistic autonomy if the dials are turned back to when he is a 20-year-old. If he is reverted to his 20-year-old evaluative scheme, the positive historicist can just treat this like any ordinary case of re-emergent autonomy. But there are two problems with this. First, this really just gets restorative autonomy by fiat, and looks like an *ad hoc* move lacking independent rationale. Second, and more worryingly, it allows a critic to start asking pesky questions about what exactly “most recent” means. If Chuck is reverted to his former self a nanosecond, ten seconds, or a minute before the manipulation, then perhaps this is kosher and his autonomy restores. But what about an hour, a week, or a year? The problem is not that this is a difficult question or is one that has a vague or imprecise answer. The problem is that it seems like the addenda to positive historicism only allow it to give binary answers (either to treat as restorative or as re-emergent) to what far more plausibly is understood as a non-binary problem. Chuck’s autonomy restores *more* quickly when he is reverted to his immediate pre-manipulation self than his self one month prior to manipulation, which restores (or re-emerges) *more*

quickly than his 20-year-old self. But positive historicism cannot account for this gradient phenomenon.¹⁰⁴

On the other hand, operativity is *exactly* the sort of factor that can explain this. While you can alter and undo the robustness of evaluative schemes with sophisticated enough technology, you can never undo the fact that something was done—you can't change the fact that some evaluative element was exercised and made more operative. Operativity by stipulation resists undoing, because it tracks strictly the historical progression of evaluative elements rather than anything structural about them. Because of this, the operant historicist can plausibly account for Chuck's restorative autonomy. We can imagine that when, at 30, he is manipulated, his murderous pro-attitude has an operative value of 200, then is replaced by an altruistic pro-attitude with an operative value of 0. Had he continued to engage it through practical activity, he might have eventually come to autonomously possess that altruistic element, but the neuroscientists undo the manipulation a nanosecond later, and he regains the murderous element. Neither manipulation changes the fact that Chuck has historically exercised that murderous element (indeed, no manipulation could change that fact), and so Chuck inherits his pre-manipulation evaluative element with the same operativity that it had. Since he does not need to start over, autonomy is restored instantaneously.

By contrast, if the neuroscientists dial back his evaluative scheme to what it was when he was 20, the operant historicist can explain Chuck's relative non-responsibility by using the same explanation he does for ordinary radical reversal manipulation: Chuck's evaluative elements at age 20 are not sufficiently operative for responsibility, and so he must engage them for autonomy to re-emerge. Since the further back in Chuck's past the neuroscientists restore his former evaluative schemes, the less operative the elements in those schemes generally will be, the operant historicist *can*

¹⁰⁴ Nor can the positive historicist appeal to some "degree of opportunity," since one could have from ages 10 to 20 many opportunities to reflect on that previous evaluative element, and only recently up to age 30 had little (but sufficient) opportunity to reflect on the element that is finagled.

account for why Chuck's autonomy generally restores more quickly when he is reverted to his immediate pre-manipulation self than his self one month prior to manipulation, which restores (or re-emerges) more quickly than his 20-year-old self.

As I flagged, this provides further evidence for the flexible threshold view of operativity. Why? By stipulation, Chuck became morally responsible when he was 10, and freely acted from the relevant evaluative element when he was 20, and so however operative the relevant evaluative element was, we know its operativity was above the threshold for free action. However, when Chuck at 30 is manipulated to become his former 20-year-old self, his equally robust and equally operative evaluative element is no longer sufficient for his responsibility, and as we discussed, he must undergo *more practical activity* for his autonomy to re-emerge. But this implies that *the operativity threshold has changed* and is thus contingent on manipulation, and this is only compatible with the flexible threshold rather than fixed threshold view. Suppose Chuck's evaluative element at 20 was first authentic at age 15 when it passed the operativity threshold of 100. At 20 its operativity value is 150, and the comparable element that is replaced at age 30 has a value of 200. Even though Chuck doesn't need to start over and his atavized evaluative element has an operativity value of 150 which is *above* the original threshold, he now needs to exercise it until it reaches a higher value, say 200 (when it becomes more operative than the element it replaces).

The advantages of a flexible threshold view of operant historicism come out further when we return to Odysseus-style cases. Suppose that a man has an O-unsheddable value of stealing money from people's wallets when he knows he can get away with it, and that the value is authentic and he has freely acted on it numerous times throughout his life. After some friends intervene and make him realize the gravity of his faults, he decides to try to become a better person, but simple resolutions are not working (after all, being O-unsheddable, it is quite fixed in his scheme). He recruits neuroscientists to strip what has been an O-unsheddable value of stealing and replace it with what will always be an

O-unsheddable value of not stealing. Had the neuroscientists done this covertly and without his consent, such an intervention would be a canonical responsibility-undermining one. But it is commonly held that people can arrange their own “manipulation” in order to not subvert but rather *promote* their own autonomy—consider someone who enrolls in hypnosis to rid himself of a smoking addiction. This is why people like Mele build conditions into their accounts that make exceptions for otherwise responsibility-subverting interventions if such interventions are arranged by or consented to by the agents who undergo them.

However, while such conditions do well in preserving autonomy when agents transition *into* those states, they struggle to account for losses to autonomy when agents transition *out* of those states. Suppose some time after the operation, while the man autonomously and repeatedly does not steal when given the chance to do so risk-free, a sadistic god creates an instant agent that is a non-historical duplicate of the man in another world, and then for both replaces the altruistic value with his original pickpocketing one. The two then live out identical lives. It seems very plausible that the instant agent becomes morally responsible far sooner than the man does. The instant agent might very well have been the one in *Impulsive-God-2* who existed only a nanosecond before having his evaluative scheme changed, whereas the man went through *extraordinary measures to avoid compulsive theft*. Interestingly, the man’s pickpocketing evaluative element is even more operative than the instant agent’s, since in his history he has exercised an identical value whereas the instant agent never has. Thus, even if one has the intuition that the man and the instant agent come to autonomously possess the pickpocketing value at the same time, this would *still* support flexible threshold historicism since he has a “head start” on practically exercising the value over the instant agent, and so since they undergo the same practical activity his threshold for autonomous possession of the value must be higher. If the fixed threshold

view were correct, then the instant agent would have to become responsible *before* the man, which is implausible.¹⁰⁵

Agents who go through great measures to avoid a certain lifestyle (or avoid being manipulated into a certain lifestyle) have their responsibility more undermined when they are then manipulated into those lifestyles than agents who haven't.¹⁰⁶ However, while the structural properties of their evaluative schemes can explain why such great measures preserve autonomy, they are not always reflected in the structural properties of their new evaluative schemes. That you arrange your own manipulation might explain why your new evaluative scheme is authentic, but it doesn't show up anywhere in that evaluative scheme. Only operativity can do this sort of accounting and keep a tally of an agent's reasons that he historically has but which do not manifest in the structural robustness of his current evaluative scheme—and only on a flexible threshold view.

5.6 Fine Tuning Operativity

I have spent most of this chapter so far articulating operant historicism as a compelling alternative to the positive and negative historicisms we critiqued in the earlier chapters of this thesis, culminating with operant historicism's ability to account for re-emergent autonomy and the responsibility relevant effects of actions during radical reversal after radical restoration.

But operant varieties of historicism are not without their problems. The *way* in which operant historicism accounts for re-emergent autonomy and tally's practical activity during radical reversal opens it up to a severe criticism. Recall first that plausible accounts of moral responsibility have *epistemic* conditions, to account for nonculpability in situations like:

Deadly CPR: Violet, who has been CPR certified, sees a man on the sidewalk collapse and the situation matches one that in her training requires immediate CPR. She begins administering chest

¹⁰⁵ The only other explanation would be that operativity resets when a qualitatively identical value is reinstated. However, I have already criticized this view.

¹⁰⁶ For further evidence, imagine *One Good Quarter Century*. If Chuck, at age 30 is manipulated into being good and remains that way for a period just as long as he has been alive, it seems implausible that when the manipulation is undone and he returns to his younger evaluative scheme everything is back to normal, and what he has done in this thirty year period has *no bearing* on his responsibility.

compressions, but unbeknownst to her, the man has a congenitally thin ribcage, which she breaks, leading to his death.

Had the man lacked the congenital bone condition, CPR would have been the obligatory thing to do and would have saved his life (for which Violet would have been praiseworthy), but Violet was nonculpably ignorant of his condition, so she is not blameworthy.

Now, suppose that some O-unsheddable value of Violet's (perhaps the desire to save another in need) is in the causal etiology of her chest compressions. While the exercise of this evaluative element does not lead to behavior for which there is moral responsibility, so operant historicists must also ask the further question of *whether the evaluative element is any more operative*. While epistemic conditions seem to provide a compelling reason for disqualifying the act as one for which Violet is blameworthy, it is difficult to see why Violet's ignorance should make the exercising of that evaluative element any less operative than if the man had lacked the bone condition. She has, after all, still acted on her desire to save another's life—the evaluative element was still exercised, and when it comes an evaluative element's operative value, that seems to be all that matters.

That a person can act without responsibility yet still increase the operative value of some of his evaluative elements should not be a surprising result; after all, manipulated agents who act from newly finagled evaluative elements do not act freely, but those elements must become more operative in order for re-emergent autonomy to be possible in the first place. However, this still has major implications for what someone can be responsible for, given that, according to operant historicism, responsibility ascriptions are a function of an evaluative element's operative value. We can build up to a powerful objection to operant historicism by first making a bad objection. The bad objection begins with the (correct) observation that just as certain epistemic constraints that determine whether an evaluative element is exercised freely don't apply to whether it becomes more operative when it is

exercised, such epistemic constraints also don't apply for many modifications to evaluative schemes.¹⁰⁷

The objection continues by having us imagine someone like:

Hallucinating Tom: Tom has a generally altruistic character and believes stealing is morally terrible. However, his evaluative scheme is such that certain extremely rare experiences would lead to him authentically modifying his evaluative scheme to contain a desire to steal money. Nefarious neuroscientists zap his brain such that he hallucinates he is in those very experiences, and he develops the desire to steal. The neuroscientists then zap his brain to make him think he is in repeated instances where he has the opportunity to steal and get away with it, and he does. They do this until the operative value of the evaluative element associated with the desire reaches the threshold for free action.¹⁰⁸ Tom then, in real life, faces an opportunity to steal and does.

For most compatibilists, this is a bullet so easy to bite it could hardly be considered bullet-biting. It is certainly unfortunate that Tom would have had far different experiences had the neuroscientists not been in the picture, and certainly the neuroscientists are to some extent blameworthy for the way he turns out. But had Tom actually gone through the events rather than hallucinate them, the ark of modifications to his evaluative scheme would be exactly the sort of thing most compatibilist accounts deem freedom-preserving, and I have conceded that there is no relevant difference, so long as modifications pass through one's rational capacities, to whether the experiences are veridical. Interestingly enough, the thought experiment changes from totally innocuous to seemingly decisive if we change one small detail: imagine that the changes to Tom's evaluative scheme do not come from acceptable modifications that go through it, but rather that the neuroscientists completely eradicate his old altruistic values and replace them with selfish and norm-breaking ones, and *then* make him hallucinate many scenarios in which such values become evermore operative. The first stage in the manipulation is clearly freedom undermining, as an evaluative scheme is bypassed and its operative

¹⁰⁷ For instance, if I read Singer's argument in "Famine, Affluence, and Morality" and this persuades me to donate my money, if I am praiseworthy for that altruism I am just as praiseworthy if altruistic neuroscientists zap my brain such that I hallucinate that I am reading the same argument. Or, if I am deprived of food and feel starvation for the first time and this makes me give greater weight to the suffering of strangers, assuming this is a germane change to my evaluative scheme, it would be just as authentic if altruistic neuroscientists zap my brain such that I experience starvation for the first time. As long as the changes go *through* my rational capacities, they do not have to be spurred by veridical experiences.

¹⁰⁸ Since the changes to Tom's scheme are by stipulation ones that go through his evaluative scheme, the desire to steal is not one that bypasses his rational capacities.

elements are replaced with inoperative ones. However, in the second stage, those inoperative elements are made to become sufficiently operative.

We can express this worry about operant historicism in a more general way. Any operant historicism provides some function that takes as an input a bypassed element's previous operative value and generates the operative value that the new evaluative element must reach in order to be authentic. Thus, for any operant historicism, there is a certain type of manipulation that does not subvert responsibility. In the first stage, radical reversal, the person's evaluative scheme is bypassed and certain elements are replaced with seriously opposing ones that now have operative values of zero. In the second stage, habituation, the person is subjected to repeated hallucinatory circumstances such that he exercises those new evaluative elements until he has undergone sufficient practical activity for them to be operative. Since the operative value of an element just depends on its being exercised, the agent's perceptions of the experience need not be veridical, meaning not only that they can be hallucinations, but also that the manipulators can induce massive time dilation in the agent and complete the habituation period over the course of only minutes (which, for the agent, could feel like years). Thus, there is a way of telling the story of manipulated Beth such that she would be responsible the next day when she wakes up as Ann's value twin (she dreams that night, which she experiences as lasting for years, and during which she is constantly in situations where she exercises her philosophy-loving evaluative elements). Call this type of manipulation Reverse & Habituate manipulation. It seems any operant historicist is committed to the preservation of autonomy in such types of manipulation.

I do not believe the operant historicist can plausibly bite this bullet. The justification for any historical compatibilist account in the first place was that one is merely caused to be the way one is through certain histories, while others are compelled, and the idea of compulsion always gets cached out one way or another by changes that go *through* an agent, or that *bypass* him. The types of bypassing manipulation that completely erase and replace one's evaluative scheme, which would be freedom-

subverting on any other historical account, are actually freedom-permissive by the lights of operant historicism if the manipulation is paired with a deliberate and prolonged exposure to circumstances that make the new evaluative elements more operative. The operativity property, which makes a historical account plausible by enabling it to account for re-emergent autonomy and avoid counterintuitive commitments to instant agents being free but not instant agents who are manipulated after a nanosecond of existence, is the same property that makes the historical account deem responsibility-preserving manipulations that clearly seem responsibility-subversive.

At this point, the historicist might suggest that we have placed too much value on re-emergent autonomy. While Mele may have a point that after enough time, manipulation that initially impairs responsibility might eventually not, if the cost is the possibility of Reverse & Habituate manipulation, the possibility of re-emergent autonomy might well be abandoned. On this view, if an agent authentically possesses some evaluative element, if that element is changed in a way that bypasses his evaluative scheme, then he is permanently inculpable for actions issuing from that element. However, this view seems incorrect in light of cases like:

Axe and AI: Suppose that humans normally become free and morally responsible agents at around age eight. Axe is an ordinary eight year old who developed an authentic initial evaluative scheme at eight, and has been freely acting for a few months. He is subjected to global manipulation that completely changes his evaluative scheme, replacing many of his O-unsheddable values with different ones that will remain O-unsheddable for the rest of his life. However, in the world in which Axe lives, advances in artificial intelligence have given nanotechnology the capacity to repair and replace human body parts with atom level precision, allowing humans to live indefinitely barring adverse external circumstances.

Abandoning the possibility for re-emergent autonomy forces us to say that Axe, after freely acting for only one year, is permanently nonresponsible, even if he lives for another 100,000 years. This seems to place too great weight on one's initial character endowment and circumstances, and does not appear to be a plausible account of moral responsibility.

Now, there is one final reply available to the operant historicist. The objection I initially raised depends on finagled evaluative elements always having the opportunity to become authentic should

they become operative enough—it just takes more and more practical activity the more operative they were before manipulation—and the reply I considered was that agents, once finagled, can never be responsible. However, the historicist can introduce lexicality into his function; the more operative an original evaluative element was, the more operative the finagled element must be to be authentic, but at some point there is some threshold of operativity of the original evaluative element such that *no* operative value of the finagled element, no matter how high, can make that element authentic. If the operant historicist makes this move, then you can't be changed to be the value twin of anyone and become responsible after a hallucinatory habituation period; some of your evaluative elements, because you have exercised them so much, are so deeply ingrained in your identity that to alter them would permanently impair your responsibility for actions issuing from them:

The Lexical Claim: There is an operative value of some evaluative element, L, such that if that evaluative element is replaced with a completely opposing one, for any operative value of that new element, even an infinite one, that element would still be inauthentic.

For the same reasons that apply when lexical claims are appealed to elsewhere, the Lexical Claim is hard to believe. It is hard to believe that there is some operative value arbitrarily smaller than L for which some operative value of a finagled element would be authenticity-preserving, but for an operative value only slightly higher, at L, there could not be some large enough operative value that would be authenticity-preserving. This would mean that if someone had been a decently good person and never stolen, if they had a desire to steal finagled into their evaluative scheme they could, after sufficient practical activity, be responsible for the stealing; but, if they had been *just a little more good*, and perhaps had *one more good thought*, no amount of practical activity could make them responsible for future stealing. The break in the continuum would seem to be because the ownership of a value of after sufficient practical activity is *qualitatively different* than less ownership, but it seems implausible to think that our ownership of our values works in this way.

In other areas of philosophy, some try to preserve the spirit of lexical views by understanding

lexicity in an *imprecise* way.¹⁰⁹ On this view, we “fuzzy up” the sharp break between operativity of a value that can be replaced with the possibility of re-emergent autonomy compared to a slightly more operative value that when replaced is permanently inauthentic. But while this seems to lessen the force of one of the problems with lexicity about autonomy—namely some sharp break in the possibility of re-emergence across a miniscule amount of additional practical activity—even imprecise lexicity implies that for some operative value sufficiently large no amount of practical activity could make a finagled value authentic. And *if* one allows for autonomy to re-emerge for less operative evaluative elements that are finagled, I am at a loss seeing why some larger operative level, even if it is far far larger, would imply a radically different kind of ownership such that re-emergent autonomy is impossible.

We appear to have reached a trilemma. I have argued that any plausible historical account of moral responsibility should cohere with three constraints:

1. It can account for at least some instances of Reverse & Habituate manipulation undermining moral responsibility.
2. It allows for, in at least some instances, the possibility of re-emergent autonomy.
3. It does not appeal to lexicity in accounting for the authenticity of one’s values.

Unfortunately, it does not appear that any historicist account can do all three. Even bracketing the objection of arbitrary differences, negative historicism satisfies (1) and (3) but does not allow for (2).

The same is true for positive historicism.¹¹⁰

On the other hand, operant historicism can account for (2) by specifying a function that takes

¹⁰⁹ See Parfit (2016) and Chang (2016) for the use of imprecision (or, in Chang’s case, parity) to build on traditional lexical views, in the domain of population ethics.

¹¹⁰ We could, in principle, revise positive historicism such that the opportunity to critically assess values does not depend on abilities one already has prior to manipulation, but could itself come from having sufficient opportunity to assess from any critical capacities, regardless of their etiology. In this case, positive historicism *could* account for re-emergent autonomy after sufficient time to reflect with one’s new values. But then the same criticism returns that positive historicism is no more plausible than some nonhistoricist version, since just as we could manipulate a person into nonhistorically identifying with whatever actions we’d like in the moment, we could manipulate a person into historically (over some period of time) reflecting on and identifying with whatever actions we’d like. And the latter seems no less freedom-preserving than the former, so it would be unprincipled to disregard nonhistoricism because one finds the former implausible, only to appeal to the latter in defending positive historicism.

the previous operative value of an evaluative element that has been changed in a way that bypasses an evaluative scheme, and outputs the operative value the newly finagled element must reach in order to be authentic. However, any operant account that does this, if it wants to satisfy (1), must appeal to lexicality and thus violate (3). On the other hand, if no lexicality is appealed to in order to satisfy (3), then every instance of Bypass & Habituate manipulation with sufficient habituation will preserve autonomy, and thus (1) will be violated. The operant historicist must provide an explanation for how he can avoid the trilemma, or if he cannot, why biting the bullet is not so problematic.

PART II: The Diachronic Conditions of Moral Responsibility

CHAPTER VI: The Other Side of History

6.1 Blame over Time

Klaus Barbie was born in Germany, the son of schoolteachers in the Rhineland. After becoming a member of the Hitler Youth, he went on to join the SS and the Gestapo. In the fall of 1942, when he was 29 years old, he was assigned as the head of the Gestapo in Lyon, where he personally tortured both adult and child prisoners, and was directly responsible for the deportation, torture, or murder of more than 14,000 Jews and members of the resistance. These acts not only earned him the infamous title of the “Butcher of Lyon,” but also two death sentences by the military tribunal in Lyon, where he was tried *in absentia* after eluding capture following the fall of Nazi Germany.

Barbie had managed to hide and begin a new life in South America under the name Klaus Altmann, until he was eventually found and returned to the French. On July 4th, 1987, the Court of Assizes in Lyon convicted Barbie of seventeen counts of crimes against humanity, affirming 340 special verdicts and negating only one—the one on whether any extenuating circumstances existed. By the end of his trial, Barbie was 74 years old, convicted of heinous acts committing over 44 years earlier. Did anything significant cut across this 44-year-long connection? During the trial, both press and prosecution took great effort to portray the defendant as still the cruel Nazi responsible for such atrocities. But Barbie, presumably from legal advice, decided to absent himself, perhaps to frustrate such efforts to identify him with that cruel historical figure being portrayed, and a strand of thought in the public discourse was that it no longer made sense to blame such an old man for the atrocities in question.

Up to this point in the thesis, we have only considered the conditions or blameworthiness *at the time of action*, and whether what has occurred in one’s history is one of those conditions. However, there is another way that history might modulate moral responsibility, and that is the question of how *blameworthiness extends over time*, or how responsibility *transfers* into the future—how what an agent did

in the past relates to his, or anyone's, responsibility in the present. This area of inquiry includes all the considerations that bear on whether a person who freely does some action is responsible for that action *at some point in the future* (and, more broadly, how someone can retain or inherit responsibility for something that was done in the past). Let us call whatever conditions need to be satisfied for an agent to be morally responsible at the time of a direct free action *synchronic conditions* of moral responsibility. But suppose Klaus Barbie, throughout the 1940s, satisfied all the synchronic conditions for moral responsibility, and so was blameworthy when he committed mass atrocities. Even if we know the correct account of synchronic conditions of responsibility, this does not tell us how to assess Barbie's responsibility for the prior direct free act at some later point in the future. Given that he was blameworthy for such actions in 1942, is the man standing before the Court of Assizes blameworthy for those actions in 1987? If someone was blameworthy at t_1 for some action A, let us call whatever conditions need to be satisfied for someone at t_2 to be blameworthy for A the *diachronic conditions* of moral responsibility.¹¹¹ This part of the thesis (Part II) will largely be an exploration of the diachronic conditions of moral responsibility, while the following part will touch on how the synchronic and diachronic conditions are related to each other (Part III).

6.2 Blameworthiness and Blame

Since blameworthiness means worthy of blame, it should be uncontroversial that there is some sense in which we should blame the blameworthy. If I have not blamed someone who is blameworthy, then

¹¹¹ There is a slight terminological clarification that needs to be made. I am deliberately avoiding a similar construction used by Matheson (2014), who distinguishes between *synchronic ownership* and *diachronic ownership*, and Khoury (2013), who distinguishes between *synchronic responsibility* and *diachronic responsibility*, and Khoury and Matheson (2018) who follow Khoury's earlier work, focusing on the blameworthy side of the spectrum and distinguishing *synchronic blameworthiness* and *diachronic blameworthiness*. These authors essentially refer to the same phenomena I will be discussing, but I prefer a construction that does not make it sound like there are *two types* of responsibility or two types of blameworthiness in this sense, and then assessing how they are related to one another. Conceptually, there is just blameworthiness, and the problem is really just whether a person who is blameworthy at some point retains that blameworthiness in the future—the question of whether people can stop being blameworthy for things over time. This seems most plausibly construed as the claim that there is one concept of responsibility or blameworthiness, and wondering what *conditions* (including time or events separating when an agent is being appraised and when the action was done) affect it.

there is a sense in which I have not given them *what they deserve*, and have made some sort of mistake by not adhering to this desert claim. But this is plausibly construed as the claim that blameworthiness provides a *pro tanto*, rather than decisive, reason to blame someone.

Because of this, it is crucial to flag that I am focused here on the conditions of blameworthiness over time (whether someone deserves blame), not those of blame (whether we should blame that person). Blameworthiness is clearly not sufficient for blame; even if someone is blameworthy for some action, I might not have standing to blame him if I have done the same act (and so my potential hypocrisy subverts the appropriateness of blaming him)¹¹² or if I am not related to him in the right way. Nor is blameworthiness even necessary: overt blame often serves pedagogical or deterrent functions, explaining why it is appropriate for parents to blame children sometimes even when the children are not responsible for such actions.¹¹³ Likewise, the Klaus Barbie trial was in many ways seen as a way of making right of the expression “Never Forget” and so was a “means by which the memory of the Holocaust was transmitted to the French public.”¹¹⁴ We might believe that even if Barbie were no longer blameworthy for his past actions, the significant cultural ramifications of expressing condemnation of the Nazi project was important enough to still merit the trial.

It is important to be mindful of the appropriateness of blameworthiness and the appropriateness of all-things-considered blame with respect to diachronic conditions since time might influence our intuitions about one but not the other. For instance, there seem to be normative reasons external to blameworthiness on the value of not fixating on the past; as beings with only a finite capacity for moral appraisals and responsibility evaluations, we might not wish to devote precious moral energy to bygone events. A person might indeed always deserve blame for some action that

¹¹² Consider, for instance, Cohen (2006) on who has standing to blame terrorists.

¹¹³ There is a valid question of whether, in the case considered, the parents are really “blaming” their children. However, I do not explore this part of the blame dialectic in this thesis.

¹¹⁴ Ledoux (1987), 41.

they have done, but there is an opportunity cost to constantly devoting my attention to this at the expense of other important tasks I could be pursuing with my moral attention.

Likewise, one function of blame seems to be to help us attain an accurate impression of someone by calling our attention to one particularly salient aspect of the person's character. This is especially useful for recent acts, where new information might help us update our impression of someone's overall character. If I have always thought you were a straight arrow and you do something morally wrong, I blame you in part to focus my attention on this previously underappreciated aspect of your character, and update my overall (emotional) impression. But if I spend all my time blaming you and harboring that negative emotional attention, I will probably have more negative attitudes toward you than I should, which is a moral error.¹¹⁵ Moreover, in the time between when you did the act for which you are blameworthy and the current moment, you have undoubtedly done myriad acts also deserving of appraisal, perhaps ones even deserving of praise. By only focusing on that one wrongful action, I fail to recognize how it meshes with the rest of your behavior, and how your intervening actions might crowd out the act in question.

All of these might be appealed to as reasons to not *blame* those for what was done long ago, but bear little on whether they remain *blameworthy* for such acts. And this is because even if blameworthiness always provides a *pro tanto* reason to blame someone, there are other considerations in reaching the verdict that one should blame him, and those countervailing considerations plausibly build up more and more over time.

6.3 The Link to Personal Identity

Klaus Barbie, let us stipulate, was morally responsible for committing act A during his time in Lyon (he satisfied the synchronic conditions for morally responsibility with respect to act A on whatever

¹¹⁵ The American criminal justice system, which has some tendency to reduce people to the worst thing they've ever done, arguably manifests this mistake.

avored account of the sufficient conditions for responsibility we like). It is extremely natural to think that, if Barbie was blameworthy when he committed A, he was blameworthy for A the next second, the next hour, and the next day. And the same would seem true for subsequent days, and weeks. It does not seem like blameworthiness dissipates over time—Klaus Barbie is still blameworthy for A.

Interestingly, there seems to be an implicit presumption about personal identity in the phrasing of this point. When we ask whether Klaus Barbie is still blameworthy for A, it appears that we are presuming that the seventy-year-old in front of us *is* Klaus Barbie. For an attribution of moral responsibility to be appropriate decades later, Barbie must still exist, so that the person decades later is the same person who did war crimes, where the “is” is the “is” of identity. Indeed, the first step in Barbie’s prosecution was the Bolivian supreme court ruling that the man in question was the same person who committed the war atrocities forty-years before, and on that basis was held criminally liable. It seems obvious that personal identity is a necessary component of the diachronic conditions for preserving blame for some synchronically blameworthy act: how can *I* still be responsible for an act that wasn’t even committed by *me*? It appears that I can only be morally responsible for my own actions, and my own actions are only those done by me. As Vinit Haksar notes, “Our system of moral and criminal responsibility presupposes that the person to be blamed or punished must be the very same individual that performed the past action.”¹¹⁶ Van Inwagen claims that:¹¹⁷

If moral responsibility is real, there must be real identity across time. If I say to my father’s second wife, ‘I hold you responsible for my father’s death’, then the person I am addressing, the person denominated by my use of the word ‘you’, must have existed in the past; she must be identical with the person who persuaded my father to forsake conventional medical treatment.

Nor is this belief unshared by the majority of philosophers working on the topic. David Shoemaker calls this thought—that responsibility presupposes personal identity—a slogan since it is extremely

¹¹⁶ Haksar (1991), 244.

¹¹⁷ Van Inwagen (2000), 15.

pervasive and almost universally assumed (and rarely defended) in work on responsibility over time.¹¹⁸

This suggests:

Necessity of PI: Person B at t_2 is morally responsible for act X committed by person A at t_1 only if B is the same person as A.

When we ask whether the old man in front of the court is Klaus Barbie, we might be, in addition to presuming that responsibility requires personal identity, presuming something else. We might suppose that when we ask whether the man before us is Klaus Barbie, if he is, then this settles the question of whether he is blameworthy for the act. On this view:

Sufficiency of PI: If person A at t_1 is morally responsible for act X, and person B is the same person A, then person B at t_2 is morally responsible for act X.

Combining these two gives us the theory that the diachronic conditions for responsibility are essentially the synchronic conditions plus personal identity; given that someone *is* blameworthy for some action, a future person will also be blameworthy for that action if and only if the two are numerically identical:

Necessity and Sufficiency of PI: Person B at t_2 is morally responsible for act X for which person A is morally responsible for committing at t_1 if and only if B is the same person as A.

Personal identity is a type of *numerical identity*, meaning that it is non-scalar and is a transitive relation.

Thus, as Matheson and Khoury suggest, we can see why presuming the necessity and sufficiency of personal identity leads to the belief that blameworthiness is interminable:¹¹⁹

It's easy to see why it would seem plausible that blameworthiness doesn't diminish across time if one thought that personal identity was sufficient for diachronic blameworthiness. Because personal identity is a transitive relation that either holds completely or not at all, it would be natural to think that diachronic blameworthiness is transitive and either holds completely or not at all. And this is just to say that blameworthiness is forever, so long as personal identity holds.

¹¹⁸ Shoemaker suggests that the following “accept or advocate” the slogan: Butler (1736, 104), DeGrazia (2005, 88–89), Glannon (1998, 231), Haksar (1980, 111), Locke (1975), Madell (1981, 116), Parfit (1984, 323–326; 1986, 837–843), Reid (1785, 116–117), Sider (2001, 4, 143, 203–204), Schectman (1996, 14). We can add to the list: van Inwagen (2000, 15), Mellor (1980, 106), Haji (1998), and Mele (1995). Important to note is that this is spread across incompatibilists and compatibilists, historicists and non-historicists alike.

¹¹⁹ Khoury and Matheson (2018), 209.

While my interests in this chapter are to explore the relationship between personal identity and moral responsibility, it should be clear now why considerations around the interminability of blame are so relevant; personal identity is commonly understood as the only constituent of the diachronic condition that, when the synchronic condition for blameworthiness has been met, is sufficient for a person's blameworthiness for a previously committed act. But some authors dispute both that personal identity is necessary for moral responsibility and that it is sufficient in this sense for responsibility. It is these accounts to which we now turn.

CHAPTER VII: The Sufficiency of Personal Identity

7.1 Theories of Personal Identity

Resolving the correct view about personal identity far exceeds the scope of this thesis, but fortunately many of the arguments I will consider can be made while being agnostic on the correct account. However, we will briefly need to rehearse the dominant views about personal identity in anticipation of several worries about connecting the diachronic conditions of moral responsibility to personal identity.

There are many theories about what kind of entity we essentially are, but the two dominant camps demarcate it by either biological or psychological criteria. According to the biological criterion, a person is the same individual as a person at some other time if and only if there is some biological continuity between the two. On the most popular version of this, animalism, we are essentially (and therefore numerically identical to) a human organism. A person continues to exist so long as the biological organism that he essentially is continues to exist, and in the same way that one's identity would persist if he lost one of his fingers, so too would one's identity persist if he lost his cerebrum but the biological organism remained intact.

Views according to which identity persists in virtue of some sort of physical continuity stand opposed to those asserting that the conditions of identity are essentially psychological, not biological. On psychological views, a person is the same individual as a person at some other time only if there is some psychological continuity between the two, where psychological continuity amounts to a sufficient number of overlapping psychological connections (memories, longstanding goals and desires, etc...). We essentially are our minds, not our bodies, and it is the continuity of minds that makes for identity over time. Importantly, the psychological view of personal identity is consistent with believing that a certain sort of biological continuity—continuity of our brains—is necessary for

responsibility.¹²⁰ This is because the structural and functional continuity of our brains may very well be necessary and sufficient for the right sort of psychological continuity—and the psychological view of personal identity is identifying the relevant *criterion* for personal identity, not the states of affairs that guarantee that that criterion is satisfied.¹²¹

I provide an overview of these because the purpose of this chapter is to explore arguments that reject the view that personal identity is sufficient for blameworthiness. These take the form of presenting agents who were morally responsible for committing some action in the past and who are numerically identical to that agent in the present, but are putatively not blameworthy for those past actions. I will argue that these particular arguments do not provide compelling cases against the view that personal identity is sufficient for blameworthiness, but remain agnostic as to whether personal identity is sufficient for blameworthiness, as the claim is secondary to my main goals in this part of the thesis.

7.2 Incapacitated Agents

Matheson and Khoury believe that people who have been severely incapacitated provide evidence that the mere preservation of personal identity is not sufficient for blameworthiness of an action for which one was morally responsible when he committed it. They have us imagine a man, Brian, who is blameworthy for an act he commits at t_1 . However, at some point in the future, he is in a car accident that results in severe brain damage, but still keeps him biologically and psychologically continuous with his t_1 self (so on both dominant views of personal identity, he remains the same person following the car accidents). However, the brain damage *is* severe enough to strip him of his moral agency (perhaps he can no longer respond to moral reasons, or some other synchronic condition for responsibility). They write:¹²²

¹²⁰ For example, see McMahan's (2002, Chapter 1) account.

¹²¹ Campbell and McMahan (2010), 285-286.

¹²² Khoury and Matheson (2018), 213.

And this would be so *even if* the brain damage that he suffered was severe enough to permanently undermine his status as a moral agent, because on neither [the biological nor psychological view] does the persistence of a person entail the persistence of the capacities that ground moral agency. Thus [the sufficiency of personal identity, on either view] entails that a non-moral agent may be blameworthy. This strikes us as confused. As Michael McKenna notes:

‘A morally responsible agent...is accountable for her conduct. She is one who can be held morally responsible for what she *has done*...Most simply, one is a morally responsible agent just in case she is, *in the here and now sense*, capable of moral action for which she could be morally responsible. This is a near tautology’ (2012: 11-12; our emphasis).

Thus, a person at t can be appropriately held responsible for a past act only if she is a moral agent at t . If Brian at t_2 has the psychological capacities of a two-year-old, then he isn’t, and can’t, be blameworthy. Blameworthiness, whether synchronic or diachronic, requires that the subject be a moral agent, and the notion of a persisting person that [the biological or psychological criteria] aim to capture simply isn’t the same as the notion of a persisting moral agent.

This passage from McKenna does not seem to support the view that the blameworthiness of one’s past actions dissolves should one lose his current status as a morally responsible agent in the way Matheson and Khoury think it does. If we look at the passage in context, his description of a morally responsible agent as one who, in the present moment, is capable of being held accountable for his actions and “is accountable for her conduct” suggests that we should read “she is one who can be held morally responsible for what she has done” as “can be held responsible for what she has (just) done, or does.” Moreover, even if we read it literally, the implication would be that *if* you are a morally responsible agent in the here and now, then you can be responsible for what you have (in the past) done, not a biconditional or the converse—that being accountable for past actions requires that you are a morally responsible agent in the here and now. It seems far too inelegant and noncontextual to read it as making the claim that you cannot continue to be responsible for past actions if you are no longer a morally responsible agent.

What’s more, the authors seem to slip between blame and blameworthiness when they conclude that “a person at t can be appropriately held responsible for a past act only if she is a moral agent at t .” This is not the claim at stake when discussing whether personal identity is sufficient for moral responsibility. Harkening back to our earlier distinction, there are myriad factors beyond whether someone is *blameworthy* that bear on whether we should all things considered blame him, and

it may well be true that we can only appropriately hold responsible, or blame, someone if she is a moral agent. But even granting that, this in no way implies that the person is not blameworthy in that incapacitated state. Consider:

Temporary Coma: John is driving in a safe and respectful manner and is hit by Jason, who is driving recklessly, and Jason is, at the time of the accident, blameworthy for causing it. Both enter a coma that lasts for forty years, and wake up on the same day. The accident has not affected either of them significantly—its effects are no more than a long induced sleep. Upon waking, John blames Jason.

I take it that when they awaken, John is justified in blaming Jason. The mere difference in the duration of their unconscious states does not make the case relevantly different from whether they had only entered a coma for 10 minutes, or were completely conscious after the accident. Presumably, this is in part because Jason is still blameworthy for causing the accident. But now let us ask whether 1) it was appropriate (for others) to blame Jason while he was in the coma during those forty years, and 2) whether he was blameworthy while being in the coma.

I find it more than plausible, and am willing to concede for the sake of argument, that it was not appropriate to blame Jason while he was in the coma—expressing moral condemnation at an unconscious person hardly seems a worthwhile endeavor. But it is far from clear to me that Jason is not still *blameworthy* while he is in the coma, and this is despite the fact that he is clearly not a morally responsible agent during that unconscious period. To claim that he is not blameworthy in the intervening years would be to suggest that he goes from being blameworthy, to not blameworthy in the comatose state, and back to being blameworthy when he awakens. But blameworthiness does not seem to be the sort of thing that should fluctuate in that way—I do not suddenly lose my blameworthiness for my misdeeds every night when I fall asleep simply because I am incapable of moral action while I sleep.

Now, perhaps Matheson and Khoury will object that it is a significant feature of Brian that his moral agency is *permanently* undermined; while it may be true that Jason remains blameworthy through his coma, or that we all remain blameworthy through our sleep cycles, this is only because he or we

temporarily lose status as morally responsible agents. First, even if this move is made, it does nothing to change the fact that their earlier claim that ‘a person is blameworthy at *t* for some past act only if he is a morally responsible agent at *t*’ is false. They would need to revise it to claim that ‘a person is blameworthy at *t* for some past act only if he is a morally responsible agent at *t* *or* he could become a morally responsible agent in the future.’ But this sounds even more bizarre—why should my desert claims *now* be affected by my ability to generate desert claims in the future? Consider a murderer who is blameworthy for killing some innocent civilian. Before the cops can apprehend him, he tries to commit suicide but botches the attempt, which only manages to undermine his capacity to respond to moral reasons. But it permanently does so, at least in the absence of significant brain surgery.¹²³ Suppose this capacity is a necessary feature of the synchronic conditions for freely committing an act, and so he is no longer a morally responsible agent, who by definition will not be responsible for his future actions (perhaps showing no regret or not apologizing to the family). It still seems plausible that he is blameworthy for his past act that he committed while he was a full-blown responsible agent. Perhaps it no longer makes sense for the family to blame him on that basis since he can no longer answer to his crime, or understand what he did wrong, etc.—though even that’s not clear to me—but this is again a separate consideration from whether he is blameworthy.

7.3 Can We Modulate Blameworthiness with Different Actions?

Perhaps we cannot absolve our blameworthiness simply by losing our ability to be blameworthy in the future. But can our blameworthiness of things done in the past be affected by what we do in the future? Though not a common assertion, the claim has received some defense. French asks us to imagine that a man named Sebastian gets drunk and damages Quincy’s property:¹²⁴

Sebastian had no intention of damaging Quincy’s property, or to get drunk. After regaining sobriety and learning of his misadventures, Sebastian, who is not yet an alcoholic, subsequently and quite

¹²³ This also makes my blameworthiness now seem to depend on whether it would be possible for neuroscientists to alter my brain to restore autonomy, which seems absurd.

¹²⁴ French (1984), 499.

deliberately returns to the local pub and proceeds to get himself roaring drunk. Again we should say that Sebastian's past is known to him, at least he is aware of the fact that he got drunk on a certain occasion and destroyed Quincy's property. Yet, he embarks upon a similar course again. Though we would have excused him from moral responsibility for the damage he did on the first spree had he subsequently modified his behavior, he did not do so and...is making the crucial past behavior, not an out-of-character happenstance, but very much in character and hence something for which, as Aristotle would say, he may be held morally accountable. [This] captures (at least to some extent) the Aristotelian idea that we do not blame people for unintentionally "slightly deviating from the course of goodness", as long as they do not subsequently practice behavior that makes such deviations a matter of character.

French here seems to equivocate between two distinct cases: in one, Sebastian fails to learn the appropriate lesson from some act that he was either not blameworthy for or had significantly mitigating circumstances around; in the other Sebastian has always had some sort of recklessness in his character, and we only learn about it after the second act confirms the suspicion. It is implausible that Sebastian's continued drunkenness makes the "crucial past behavior, not an out-of-character happenstance...very much in character" in the sense that his future actions retroactively change the character dispositions he had in the past. Either his initial crash into Quincy's property was in his character or it wasn't. If it was in his character (all along), then when he does so in the future, we blame him more for his past action because we realize our first excuse was the result of an *epistemic* deficiency (of not realizing that it was in his character, and mistaking it for being out of the blue). If it wasn't in his character, then we do not blame him more for his original crash—which by stipulation in this scenario, was out of character and, as French asserts, for this reason non-culpable—but rather that he has failed to learn about the consequences of such behavior. Thus, we judge Sebastian for more than just his second reckless act either because it makes us realize an earlier epistemic failure on our part or a moral failure on his, but in neither of these cases does Sebastian's blameworthiness for crashing into Quincy's property actually modulate.

Khoury offers a similar case to argue that blameworthiness can increase over time, asking us to imagine Jane who half-heartedly insults her brother at t_1 . In one scenario, time "goes by and she no longer feels bad about it at all, but is very pleased with the insult, and if she had it to do over again

would have offered a crueler remark.”¹²⁵ In another scenario, it remains half-hearted. Khoury asserts that “most will have the intuition that Jane in the former case is more blameworthy for the insult at t_1 than she is in the latter case,”¹²⁶ but if that is true then I am in the minority. There could be two reasons we would be tempted to judge Jane differently in the two scenarios. One, that in the former she is a psychologically different person who is less similar, and so identifies less with, the bad-making features of the insult, will be the subject of the next section. The other is that her subsequent actions (of abandoning her guilt, or forming the intention to be even meaner) enhance her blameworthiness for the insult. But if the latter is appealed to, then just as in the case of Sebastian, the interlocutor who believes blameworthiness cannot be affected by future actions can more than plausibly argue that Jane’s future actions are just future wrong actions for which she is blameworthy—she is not *more* blameworthy, but she is blameworthy *for more*. This leaves only the first consideration: can radical changes in one’s psychology affects past blameworthiness?

7.4 Can We Modulate Blameworthiness with Different Psychology?

Locke asks us to consider appraisal of two consciousnesses who share a body.¹²⁷

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?

The expectation is that we judge the day consciousness separately from the night consciousness, and if we link responsibility to identity, each is responsible for and only for what he freely does when in control of the body, which I do not dispute. However, as a matter of pressuring the sufficiency claim of personal identity with respect to the diachronic conditions of responsibility, the thought experiment needs some alteration in light of psychological views of identity. In so far as each person acts according to the synchronic conditions for moral responsibility but we stand by the intuition that each is not

¹²⁵ Khoury (2013), 742.

¹²⁶ Ibid.

¹²⁷ Locke (1975), 48.

responsible for the actions of the other, this puts pressure on the sufficiency claim *if the biological criterion is invoked*. The day- and night-persons share a body and so would be numerically identical on a biological view of identity, but presumably the radical break in psychological connectiveness is below the threshold on any plausible psychological criterion. David Shoemaker offers us an updated scenario of two seemingly different people, Johann and Sebastian, sharing the same body:¹²⁸

Consider *Gradual Fugue*: suppose that [a fugue] transformation from Johann to Sebastian and back occur gradually (the product of a tumor that grows and then wanes, say) such that psychological continuity is preserved all along. Regardless of such continuity, the Sebastian-actions still do not seem to belong to Johann: while he now remembers them, he nevertheless repudiates them as the actions of someone with a very different psychological make-up and motivational set, actions that seem to Johann as those of a stranger, inexplicable in motivation and justification. Nevertheless, a psychological continuity view of ownership implies that the Sebastian-actions *do* belong to Johann. Psychological continuity is thus insufficient...

In *Gradual Fugue*, personal identity is preserved on both the biological and psychological criteria. Johann and Sebastian are both biologically and (if the tumor growth is slow enough) psychologically continuous with one another, and so let us assume that we are really looking at one person all along. I share Shoemaker's intuition about the non-responsibility of Johann for the Sebastian-actions. If the person freely makes both Sebastian-actions and Johann-actions and so is responsible for them at the times they are done, then *since* personal identity endures it looks like personal identity is *not sufficient* for responsibility over time.

But there's the rub: while I accept the conditional, I deny that the antecedent is true. Harkening back to our discussion of historicism, a common justification is that substantive changes done to one's evaluative scheme without one's desire or consent bypass it and undermine responsibility. And tumors that cause massive changes to a person such that he actually appears to be two different people inhabiting the same body are *canonical* examples of changes to a person that seem to preserve non-historicist conditions for (synchronic) moral responsibility but are nevertheless freedom-undermining.

¹²⁸ Shoemaker (2012), 114.

Thus, while Shoemaker and I agree *that* Johann is not responsible for the Sebastian-actions despite being numerically identical to Sebastian, we disagree *why*. The view that synchronic responsibility plus personal identity guarantees responsibility, paired with the judgment that Johann is not responsible, entails the disjunction of 1) he does not satisfy the synchronic conditions for responsibility when making Sebastian-actions (he does not freely Sebastian-act), 2) he is not numerically identical to Sebastian, or 3) the sufficiency view (of personal identity as the diachronic condition) is false. Shoemaker seems to assume that 1) is false, and since 2) is false, he concludes 3). However, one can easily uphold 3) by simply claiming that 1) is true, and since this is a claim that nearly every historicist would agree with, the thought experiment hardly makes the case against the sufficiency view.¹²⁹ At best, it shows an interesting connection between historicism and personal identity and how the two relate to moral responsibility. But this is an issue I will explore in greater length in the Part III of this thesis. That consideration aside, *Gradual Fugue* underscores a crucial principle that we need to bear in mind when discussing personal identity and moral responsibility. We have strong intuitions about both, and so whenever we connect the two we need to make sure what we are not confusing an intuition about one for an intuition about the other.

Khoury and Matheson offer an alternative example that they believe demonstrates the insufficiency of personal identity. They imagine an explorer, Leon, who drinks from a well that enables him to live for centuries:¹³⁰

So, Leon at age 500 is biologically and psychologically continuous with but psychologically *disconnected* to Leon at age 250; Leon at age 250 is biologically and psychologically continuous with but psychologically disconnected to Leon at age 50. Suppose that Leon at age 30 (Leon-30) – the age he drinks from the well – has the psychological profile of a ruthless conquistador. Imagine that Leon-30 commits a reprehensible act M – namely the massacre of the island’s native inhabitants. And imagine that Leon-30 satisfies the conditions on synchronic blameworthiness for doing so. So Leon-30 is synchronically blameworthy for M. Now fast-forward 500 years. Leon is now 530 years old. He has

¹²⁹ No one could plausibly think identity could ever be a sufficient condition for MR [moral responsibility] anyway—there are surely epistemic and control conditions one must meet in addition to identity—so identity can serve as no more and no less than a necessary condition for MR. To reiterate, the sufficiency condition being evaluated is whether personal identity, *conditional on the synchronic conditions obtaining*, is sufficient for moral responsibility.

¹³⁰ Khoury and Matheson (2018), 213-214.

gone through many characters during his life. Suppose that Leon-530 is a perfect time-slice psychological twin of your favorite moral saint. Leon-530 shares *no* distinctive psychological features with Leon-30 but remains psychologically and biologically continuous with him. Now, is Leon-530 blameworthy for M?

We contend that Leon-530 is *not* blameworthy for M. Why? Because Leon-530 shares none of Leon-30's *distinctive* psychological features, and these distinctive psychological features were essential to him committing M. While these two stages of Leon might be connected by the relations of psychological and biological continuity, they aren't connected by any direct distinctive psychological connections themselves. So the psychological states (e.g. beliefs, desires, or intentions) from which M stemmed aren't shared by Leon-530. Further, having the full psychological profile of your favorite moral saint, Leon-530 has no memory of the action M; there's simply no lasting trace of the "springs of action" that gave rise to M. Given this, we find it intuitively compelling to think that Leon-530 isn't blameworthy for M.

I agree that Leon-530 is numerically identical to Leon-30. Stripping away irrelevant details, the case is really just of one where humans live extended lives, and asks whether they are blameworthy much later in their lives for actions done at a time when causal springs from which such actions flowed bear now resemblance to their current springs of action. The (only) principle that the authors seem to appeal to is that if there is no trace of one's causal springs essential to some prior act in his current psychological scheme, then he is not responsible for that previous act. I believe this principle has implausible implications for who is and is not blameworthy over time.

The first thing to flag is that this principle invokes no conception of time or duration. The long passage of years might be an explanation for why his psychological elements changed so much, but it is not the criterion by which we evaluate his current blameworthiness for the massacre. This is unsurprising; as our discussion of John, Jason, and the comas after a car accident seemed to suggest, there is nothing intrinsic about the (mere) passage of clock time that relates to blame. But the passage of time often includes many events that are relevant to blame. In any case, the appeal to centuries separating Leon in the present from his earlier massacre, I believe, obscures some of the implausible consequences of invoking this principle.

First, since there is no appeal in duration contained within the principle, it applies to even the shortest of timespans. But consider:

Traumatic Conversion: Some things are just too cruel to live with. Eric kidnaps, tortures, and kills a friend, and freely does so (and so is blameworthy for his actions). However, the experience affects him far more than expected. It is so traumatizing that it leads him to completely suppress all memories of having done so, and undergo a moral conversion on par with a religion conversion. After the incident his temperament completely transforms to one qualitatively identical to a moral saint, and he has no memory of the incident.

The situation is as absurd as it would be to believe that post-conversion, Eric, in virtue of sharing no psychological elements that were essential to his heinous act, is no longer blameworthy for the crime he committed the previous day. But the principle that Matheson and Khoury suggest as explaining the innocence of Leon delivers this implausible judgment. It seems that the only reply available to the authors is one they make to the objection that one of (the descendants of) the victims of Leon's massacre could still appropriately blame him:¹³¹

We agree that it may be appropriate for a victim to blame Leon-530, but we think the sense of appropriateness is merely epistemic. If your friend has overwhelming evidence that you, contrary to fact, are guilty of spreading a vicious rumor about her, then her blame of you is epistemically appropriate; it accords with her evidence. But it doesn't accord with the facts; you aren't actually blameworthy for spreading the rumor. Similarly, given that Leon-530 is personally identical with Leon-30 this may be evidence that Leon-530 is diachronically blameworthy for M. And so the victim's blame of Leon-530 for M may be epistemically appropriate. But this doesn't entail that Leon-530 is blameworthy.

First, this seems to sidestep the heart of the objection, which I presume holds that a victim could still appropriately blame Leon *even after* learning about his psychological divergence, and hence, on their principle, his non-responsibility (and moreover that appropriately means justifiably given the scenario, not given his perspective). But in any case, the basic form of this reply applied to *Traumatic Conversion* is that when I judge that Eric is still blameworthy after his conversion, I am failing to appreciate his change in psychology and am using the generally reliable but defeasible heuristic that someone who was blameworthy for a crime yesterday is also blameworthy today. But I suspect it is more likely that the reported intuitions around Leon are confusing the appropriateness of overt blame over time with the appropriateness of blameworthiness following psychological changes.

¹³¹ Ibid, 215.

The principle is even more implausible than *Traumatic Conversion* suggests. Consider:

Arranged Conversion: Eric reads that psychological disconnectedness can absolve responsibility, and comes up with an ingenious plan. He kidnaps, tortures, and kills a friend, and freely does so (and so is blameworthy for his actions). However, he arranges with some local neuroscientists to have all the memories expunged, and for his evaluative elements that were essential to the action to be replaced with altruistic ones. After the arranged conversion his temperament completely transforms to one qualitatively identical to a moral saint, and he has no memory of the incident.

Given that Eric is the same person as he was pre-conversion, I find the evaluation that Matheson and Khoury's principle implies—that he is no longer blameworthy—incredible. It seems implausible that criminals can dodge responsibility by having the requisite evaluative elements removed from their schemes that govern practical deliberation and action. If neuroscientists one day create a button that people can press to voluntarily have their psychological features altered to be qualitatively identical to those of moral saints, there would be no blameworthiness among the people realizing they could get off scot-free after murdering their enemies or rivals simply by pressing the button. The implausibility of this implication remains even after controlling for short durations between radical psychological divergences. Suppose Leon enjoys living for centuries but is tired from constantly being blamed for a massacre he did so long ago, so Leon-530, who hasn't changed at all from the murderous Leon-30 enlists his neuroscience friend to eliminate the causal springs that issued in his massacre. This does not seem to be a germane way of dodging blame.

Obviously, the range of arguments considered so far are far from exhaustive of those that might be made against the claim that personal identity is sufficient for sustaining blameworthiness over time, and since I have only provided objections to these arguments rather than general arguments against the possibility of blameworthiness dissipating when identity is preserved, I want to emphasize that my goal here is not to definitively argue that personal identity is sufficient in the relevant sense for sustained blameworthiness. I care less about the sufficiency claim, and remain agnostic as to whether personal identity is sufficient to extend future blameworthiness versus being just a necessary feature of the diachronic condition that other factors might modulate. My aim rather stems from the

fact that many of the authors discussed propose alternatives to personal identity as necessary for the diachronic condition *because* they think its inadequacies as a sufficient condition suggest that something else is doing the heavy lifting. Since in the next chapter I will argue against these accounts that dispute personal identity as a necessary component of the diachronic condition of moral responsibility, I minimally wish to put pressure on the considerations that have motivated people to search for alternative accounts in the first place. Perhaps the best account of the diachronic condition will require personal identity, but while you are only responsible for your own free actions you might only be responsible for some of your prior free actions. My main goal in this part of the thesis is to assess that claim of whether you can only be responsible for your own free actions, and it is this that we turn to next.

CHAPTER VIII: The Necessity of Personal Identity

8.1 Pressure on the Necessity Claim

In the last chapter, I considered a variety of arguments meant to put pressure on the idea that if someone is blameworthy for an action, then in the future someone numerically identical to that blameworthy person will also be blameworthy for the action. In other words, I considered arguments against the view that blameworthiness plus personal identity is sufficient for future blameworthiness. However, many of these arguments, as flagged, were offered as a way of demonstrating that personal identity is not the diachronic condition on moral responsibility, in anticipation of some other criterion for determining responsibility over time (i.e. invoking some variation of psychological connectedness, as we explore in far greater detail in what follows). Recall that if psychological connectedness matters for moral responsibility over time, it could be relevant in two distinct ways. On the first (which disputes the *sufficiency* of personal identity), psychological connectedness modulates one's responsibility for past actions, where you might be non-responsible or less responsible for some of your acts on the basis of further facts about psychological connectedness, but you still can *only* be responsible for your own past acts. On the second (which disputes the *necessity* of personal identity), you can actually be responsible for *other people's actions*, and in this chapter, I consider such alternatives which do not even require personal identity for blameworthiness over time.

The search for such accounts often begins with the observation that there are plausibly instances where personal identity is not preserved, but moral responsibility is. Ben Matheson has us consider a variation of our Chuck/Beth manipulation cases.¹³² Recall that in one version, Chuck is a murderous villain and Beth is the sweetest person you know, and Beth has her values replaced by neuroscientists to become Chuck's value twin. In other words, Beth's values are replaced but her background psychological contents (her psychological contents that are not her values) remain the

¹³² Matheson (2014), 46-47.

same. In Matheson's version, Chuck is the one who is manipulated, but instead of being given Beth's values he is given her background psychological contents, so that post-Chuck and post-Beth (from the original thought experiment) are non-historical duplicates. I have some doubts about the coherence of this thought experiment (how will post-Chuck, who has male genitalia, reconcile this with the background belief that they are a woman named Beth?)¹³³, but I will concede them to be as charitable to the case as possible. Matheson asks:

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.¹³⁴

Before I challenge the plausibility of this claim, I want to flag the observation that, in defending post-Chuck's responsibility, and as I have italicized for emphasis, Matheson continually uses pronouns suggestive of the preservation of Chuck's identity through the manipulation. However, if we intend to offer a case as supporting the claim that personal identity is not necessary for preserved moral responsibility, it seems misleading to write it in a way that, while it technically eliminates (on the psychological criterion) personal identity, sounds like it presupposes it. Writing on a similar case where psychological continuity is disrupted, but in more accurate terms and therefore more congenial for exploring the necessity claim, Shoemaker also claims that personal identity is not necessary for

¹³³ Matheson touches on this, writing (2013, 325):

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.

But, as I have mentioned, this ignores the fact that while they are psychologically identical they are not physically identical. And they do not live in the same houses, or have the same pets, or interact on a daily basis with the same objects in their external environment, which Matheson overlooks.

¹³⁴ Ibid. 324.

preserved responsibility:

Consider *Disproportionately Disciplined Doug*: suppose that Doug gently mocks a young girl and then is beaten up so badly by her parents that the beaten-up inheritor of Doug's body in the morning is psychologically discontinuous with Doug (given that there is an insufficient number of psychological connections between him and Doug), but nevertheless suppose the beaten-up person retains a memory of mocking the child and still has a smaller number of beliefs, desires, and character traits that were nevertheless essential to his being moved to mock her. Here I suspect we would agree that the mocking-action is still attributable to the assault survivor [and so he is morally responsible for the mocking], despite the fact that he is not psychologically continuous with Doug.¹³⁵

Even if one agrees with Matheson's and Shoemaker's reasoning, the reader may point out that this only invalidates the necessity of personal identity for preserved responsibility on the *psychological criterion*, and therefore supports only the more modest claim: if the psychological criterion of personal identity is correct, then personal identity is not necessary as a diachronic condition of moral responsibility. Now, there are important features about the two cases that may prove relevant to whether responsibility is preserved or not. For instance, the inheritor of Doug's body shares physical brain continuity with Doug, as does post-Chuck and pre-Chuck. While I will more systematically assess the relevance of these features in the next two chapters, I only raise the cases now to motivate the idea that sometimes personal identity appears to not be preserved while responsibility is, on a psychological criterion of personal identity. With just this, it remains unclear whether the same would apply on the biological criterion.

But the canonical objection used against the biological criterion of personal identity can also be deployed for this purpose. Consider *Brain Transplant*, where after a man commits some crime (say he stabs a bystander) his entire brain is transplanted into another person's empty head. Presumably, the person who wakes up will be exactly like the stabber psychologically—he will remember the crime, have the values of and dispositions that prompted the crime, etc...While this is often used to put pressure on the biological criterion (since it seems like the new person will be the same person as the

¹³⁵ Shoemaker (2012), 114-115.

robber), it also seems clear that the new person will be responsible for the crime. But if this is true, then this will be in spite of the new person being a different person from the stabber, on the biological view. If the intuitions evoked from *Brain Transplant* and *Disproportionately-Disciplined Doug* are correct, then we can conclude that if either the biological or the psychological criterion for personal identity is true, then personal identity is not necessary for preserved responsibility. But if not, what else is? What account can explain the preserved responsibility in the two test cases?

8.2 The Psychological Connectedness Account Following the Parfitian¹³⁶ Tradition

To my knowledge, Matheson, Khoury, and Shoemaker are the only ones to offer explicit arguments¹³⁷ against the claim that moral responsibility over time presupposes persistence of personal identity.

¹³⁶ Shoemaker interprets Parfit (1984, 323-326) as actually endorsing the view that desert requires identity, presumably because he only makes the general claim that “When some convict is now less closely connected to himself at the time of his crime, he deserves less punishment,” in which case psychological connectedness functions as a modulator but identity is still required, rather than functioning as a substitute for identity. I believe Parfit at least considers the possibility when he writes that “We can defensibly claim that psychological continuity carries with it desert for past crimes... But I have not yet found this argument.” Nevertheless, exegetical considerations aside, Parfit certainly does not explicitly defend the view that responsibility in the sense we are discussing here does not require identity.

¹³⁷ By explicit arguments I mean both that the concept of responsibility is explicitly addressed, and that the author offers a series of considerations about the claim rather than highlighting a relevant intuition or stating *that* it is the case without arguing *why* it is the case. Regarding the former, many authors argue that *what matters* does not require personal identity, and plausibly moral responsibility is either a part of what matters or supervenes on what matters. However, widening my discussion to what matters exceeds the scope of this thesis, and so I only intend the arguments to tactically address moral responsibility, while recognizing that many could also be applied to the dialectic on what matters. Regarding the latter, many authors seem to accept that moral responsibility over time does not require personal identity, but these are not argued for at length. To mention just a couple:

- Pummer (2014) offers a series of arguments building from the idea that fission/division preserves responsibility despite not preserving identity. However, he flags that he takes this as a given to explore its implications: “It is important to notice that, in denying [the view that] Desert Requires Identity, I am only claiming that if Desert is true [and we deserve things in virtue of our wrongdoing], then my continuers might well deserve punishment for my wrongdoing, even if they are not identical to me. I am not committing to any more specific view about what it is about my continuers that makes them deserving of punishment for what I did” (53).
- Wiggins (1976, 138) writes that “a malefactor could scarcely evade responsibility by contriving his own fission.” Both Pummer and Wiggins seem to come to this from the view that in the nondivision case where my brain is divided but only *one* half is preserved, the resulting continuer would be identical to me and also share in my responsibility. But it is only a technicality stemming from the logic of numerical identity that in the division case neither continuer can be me that prevents either of the continuers from being me, and it is hard to see how a person can get off crimes “scot-free” (Pummer 2014, 52) simply because of the technicalities of the logic of identity.

Importantly, both of these are in the context of fission cases (in which there is a causal relationship between the two psychologically connected persons) and since I am only considering non-causal versions of the psychological connectedness account here, I deal with such considerations in the following chapter.

Interestingly, they all propose similar accounts, those that require some form of *psychological connectedness*. To refresh, psychological connectedness refers to holding particular direct psychological connections. The most obvious example of a connection is a shared memory of some past event, but could also include a persisting desire or value, acting on a previously formed intention, etc... This is a scalar notion, since the *number* of particular direct connections between A at t_1 and B at t_2 can vary, and intransitive, since A can be psychologically connected to B, B can be psychologically connected to C, but C need not be psychologically connected to A. Importantly, invoking an alternative account that grounds responsibility with psychological connections will either have no causal requirements for the relationship between those connections or it will have some causal requirement, and in this chapter I evaluate the plausibility of non-causal, or what I will call, *general*, psychological connectedness accounts.¹³⁸

On a general (non-causal) version of an alternative to personal identity for the diachronic condition of responsibility, whether (and how much) a person is responsible for some past free act depends on the psychological connections between the two. As Khoury and Matheson write in what I will call the general psychological connectedness account:¹³⁹

The General Psychological Connectedness Account of the Diachronic Condition: “If a person A at t_1 is blameworthy to degree d for act X which occurs at t_1 , then an individual B at t_2 is blameworthy for X to degree d if B is maximally psychologically connected to A” in the appropriate way (or maximally psychologically

¹³⁸ As a matter of book-keeping, I will follow the tradition of using the term “psychological connection” to refer only to the relationship of the intrinsic features of two psychological states, not their external features (i.e. whether they are causally related or not). In the personal identity dialectic, disagreements often arise in virtue of a disagreement on whether there are constraints on the *cause* of continuity between psychological connections. But to avoid creating new terms when I explore psychological connections with different types of causal relations to one another, I will count any two states with the right qualitative similarity as a psychological connection, and qualify (as we will see shortly) such connections as, for instance, general psychological connections or wide or real psychological connections according to their causal relationship. Obviously, if in fact there is some causal component to the actual holding of psychological connections, the same points could be made, but under a somewhat different presentation.

¹³⁹ Earlier, Khoury (2013, 741) writes: “The transfer of responsibility through time is directly sensitive to psychological connectedness. In cases in which an agent commits some act X at time t_1 because of some particular psychological structure, the agent’s responsibility at some later time t_2 is determined by the degree to which and the way in which that psychological structure is connected to the agent’s later psychology. That is, if the agent is blameworthy for some act X to degree d at time t_1 , then when there is maximal psychological connectedness between time t_1 and some later time t_2 the agent is blameworthy for X to degree d at t_1 . Most interesting here is his use of “the agent” rather than *an* agent, which makes it sound as if the account is challenging the sufficiency, rather than the necessity, of personal identity.

connected with respect to the relevant subset of psychological states).¹⁴⁰

Shoemaker endorses something similar with:¹⁴¹

Suppose that I perform some immoral action today and then while I sleep tonight I am secretly duplicated by some mad scientists who then kill me in a few days, cremate my body, and secretly insert the duplicate into my place. The duplicate, it seems clear, cannot be me. Nevertheless, it seems that my actions are still attributable to him for purposes of accountability. The pull of this intuition has its most compelling source in a vivid representation of the phenomenology of the duplicate. He will come into existence full-blown thinking he is me, and he will (quasi-)remember my actions, will delight in thinking that he's gotten away with the immorality, will carry out my intention to celebrate the immorality, and so forth. Once we imaginatively project ourselves into his mental terrain and appreciate that he is *psychologically exactly similar*¹⁴² to the *real* me, it becomes extraordinarily difficult to resist the verdict that my actions are properly attributable to him as well, that he is just as eligible for being held accountable as I am.

In both of these accounts, there is no real causal requirement for how the latter person's psychology came to be related to the former's; it is solely in virtue of the non-historical nature of each person's psychological components that the latter becomes responsible for the former's deed. (There is technically a causal relationship in Shoemaker's example since the mad scientists determine the qualitative structure of the duplicate based on the original body—his justification, however, does not appeal to this fact at all, as the justification for transferred responsibility comes from “phenomenology” and phenomenology is not a causal concept.¹⁴³ In the following chapter, when I consider causal variations of the psychological connectedness view, I will consider such nuances more systematically).

By making the diachronic condition of responsibility a function of distinctive (and relevant) psychological connectedness, this alternative account has several ostensible advantages. First, in so far as one believes that blameworthiness diminishes over time, the psychological connectedness account

¹⁴⁰ The qualification of the “appropriate way” across the “relevant subset” stems from the plausible qualification that connectedness only need obtain across the psychological elements implicated in the action, but I will note this shortly.

¹⁴¹ Shoemaker (2012), 120.

¹⁴² Emphasis mine.

¹⁴³ This passage notwithstanding, Shoemaker (ibid.) goes on to present an account of the diachronic condition that *does* require some causal constraint on psychological connections. Since my purposes here are primarily evaluative rather than expository or exegetic, I do not intend to parse exactly what view Shoemaker is actually endorsing but rather to systematically examine the strongest and most charitable versions of the range of views one could take.

can do what a personal identity account can't; since psychological connectedness is scalar, a person's responsibility in the future for something he did in the past admits of degrees, whereas the non-scalar personal identity account as a (sufficient) diachronic condition makes the blameworthiness all or nothing. Thus, in so far as one does judge Leon-530 as less blameworthy (or not blameworthy at all) for the massacre committed by Leon-30, we have a viable explanation: Leon-530 has no relevant psychological connections with his former self.

Moreover, this account can explain the putatively intuitive judgments in our test cases. Post-Chuck, albeit psychologically discontinuous with pre-Chuck, shares direct psychological connections with pre-Chuck's values and evaluative elements that led to the murder, and thus inherits the blameworthiness, as does the beaten-up person in *Disproportionately-Disciplined Doug*. Likewise, in *Brain Transplant*, the person, when he awakens, will have near complete psychological connectedness with the stabber, and for that reason inherits the responsibility for his murder.

There is a further question, as noted, whether the diachronic condition is a function of the total psychological connectedness between the past responsible party and the present person whose responsibility is in question, or whether "some subset of those connections get special weight."¹⁴⁴ For instance, Khoury argues that what primarily matters is the connectedness between the psychological elements that were implicated in the causal etiology of the earlier action and the present person's psychology, and Shoemaker similarly states that "only a certain *subset* of psychological relations is relevant to ownership,"¹⁴⁵ and then goes on to specify what exactly that subset is. However, I will not consider this nuance because I think the general account (regardless of how we answer the subset question) is vulnerable to decisive objections.

¹⁴⁴ Khoury and Matheson (2018), 219.

¹⁴⁵ Shoemaker (2012), 115.

8.3 Freak Coincidences and Post-Hoc Retributions

Let us return to post-Chuck and *Disproportionately-Disciplined Doug*. The connectedness account tells us that even though such individuals are not the same persons as the ones who committed crimes in the past, they bear such an intimate relation—that of sufficiently strong psychological connectedness—to them that they share in their responsibility, their blameworthiness. However, as I noted earlier, we need to be mindful of what exactly is evoking our intuitions. Doug, while he may not have psychological continuity, *does* have bodily continuity. In so far as our pre-theoretical conception of personal identity has mixed biological elements (or our intuitions about the relevance of bodily continuity persist even after we theoretically discredit them) and that our pre-theoretical conception of moral responsibility presupposes personal identity, then *even* if we on reflection endorse a psychological criterion, we run the risk of our intuitions about moral responsibility being influenced by a persisting view that bodily continuity matters for personal identity, and that personal identity matters for responsibility. That possibility is completely eliminated if we instead consider an example from Nozick:¹⁴⁶

As you die, a very improbable random event occurs elsewhere in the universe: the molecules come together precisely in the configuration of your brain and a very similar (but healthier) body, exhibiting complete psychological similarity to you...This is not you; though it resembles you, by hypothesis, it does not arise out of you. It is not any continuer of you.

If we wish to overlay this on the Doug example, we can borrow from Nozick's case and imagine that instead of appearing elsewhere in the universe, we have:

*Disproportionately-Disciplined Doug**: Same as *Disproportionately-Disciplined Doug*, but once the beaten-up person awakes scientists obliterate his body in its entirety, and this moment happens to coincide with the improbable configuration that leads to a near molecule-for-molecule identical person appearing in exactly the same location.

I have a far weaker intuition that this instantaneous person is responsible—but there is no difference in *Disproportionately-Disciplined Doug** with respect to the general psychological connectedness between

¹⁴⁶ Nozick (1981), 41.

the engineered person and Doug, and the original beaten-up person and Doug. Of course, Shoemaker's reasoning still applies. Our coincidental person retains the memory of insulting the kid, has the beliefs, desires, values, and principles of practical deliberation that were essential to insulting the kid, and thus is related to the act in a way that is qualitatively identical to the way that Doug was. But there seems to be a different way of telling the story: there is now a new person, who happens to have some relevant psychological connections with Doug, but he *hasn't actually done the act of insulting the kid*, he just mistakenly thinks he has. Indeed, we can contrast *Disproportionately-Disciplined Doug** with:

Coincidental Doug: Scientists devise a machine that can create people with any range of possible values, beliefs, desires, etc...They simply need to turn the dials on their machine to the relevant setting, and they can create a person who is non-historically identical to any person in history. The machine has a 'random' setting that randomly permutes all the possible combinations to create a psychological profile, and the one created coincidentally is qualitatively identical to the beaten-up body (and to the engineered body in *Disproportionately-Disciplined Doug**).

On the general psychological connectedness account, there should be no difference in the verdicts we deliver to the newly created people in *Disproportionately-Disciplined Doug** and *Coincidental Doug*. Whether accidentally created by neuroscientists or simply emerging out of a spontaneous collision and unification of a sludge of atoms through blind chance, on the connectedness account all these created persons are qualitatively identical to the beaten-up body and so share the same connectedness with Doug. But when the story is told this way, it seems like such created individuals are much more the *unlucky* inheritors of memories of crimes they haven't committed and for which they are not responsible.

Which of these two ways of understanding the newly created Doug-duplicates is correct? We can isolate the relevance of psychological connectedness by beginning with:

Fission: Doug is beaten up, resulting in the psychologically discontinuous person, call him Doug*. Doug's brain, like ours, has two hemispheres. Surgeons divide Doug's brain and transplant both hemispheres into two brainless bodies that are clones of Doug's, and cause each half to regrow its missing half, resulting in two people who are exactly like Doug*.

Both of the Doug* clones, call them Lefty and Righty, seem psychologically and biologically continuous with Doug* (and we can imagine both a biological and a psychological identity theorist upholding the preservation of identity if one of Doug*'s hemispheres is destroyed and the other is transplanted). However, since Lefty and Righty are qualitatively identical to one another, both are equally connected and continuous with Doug*. Now either 1) both Lefty and Righty are Doug*, 2) only one is, or 3) neither is.

With respect to 1), although Lefty and Righty are qualitatively identical, they are numerically distinct (they are not one and the same person). But since the personal identity relation is transitive, given that they are not identical to each other, they cannot both be identical to Doug* (another way of putting it is that the personal identity relation is one of numerical identity, and thus is one-one, but Doug*'s relation with the fission products is one-many, and thus this relation cannot be one of numerical, and therefore, personal identity). With respect to 2), since Lefty and Righty are symmetrically related to Doug*, they each seem to have equal claim to being Doug*, and so barring that we are essentially immaterial souls that randomly pop into one of the clones in such fission cases, it is impossible for one to be Doug but not the other. On any standard view of personal identity, Doug* has died and is survived by two new individuals. Shoemaker, in considering a similar case, still believes there is some preserved responsibility:

So if I had performed some immorality and then undergone fission, that pre-fission action could not be properly attributable to anyone anymore, despite the fact that *two* people would “remember” doing my deeds and relish the feeling of having gotten away with them (and they would, in addition, be exactly like me in every respect). And if attributability is necessary for accountability, neither fission product could be accountable (to any degree) for my actions. This will strike many of us as the wrong answer, however—surely, my actions are (at least partially) attributable to *both* fission products—and so the fission case seems to thwart the viability of any relationship between [moral responsibility] and numerical identical.¹⁴⁷

I wonder how much the deliberate undergoing of fission *for the purpose of evading responsibility* influences

¹⁴⁷ Shoemaker (2012), 118.

our intuitions about preserved responsibility. After all, a malefactor *could* evade responsibility by contriving his own suicide, and if fissioning himself results in *his* death, the way he scarcely evades responsibility is rather misleading. Fortunately for us, the general connectedness account does not require us to give special judgments for deliberately undertaken fissions, nor anything about the causal relationship that the original brain and its fissioned halves might have with one another. On this view, we should judge *Fission*, and deliberately undertaken fissions, symmetrically with:

Coincidental Fission-equivalent: Scientists have the same people-creating mission as in *Coincidental Doug*. It has two slots that allow them to create two people simultaneously. They set it on ‘random’ again but have it create two duplicates of the same random person. The two newly created people are identical to Lefty and Righty—call them Lefty* and Righty*.

Yes, Lefty* and Righty* both have memories of Doug’s insult and the values that led to it. But they are also not Doug, or Doug*, and have not even done the act that Doug did. Moreover, since the passage of time has no necessary connection with the psychological connectedness of things currently in existence, if we visited another planet at some point in the future and found that they did possess the technology and had randomly created Lefty* and Righty*, would we demand to hold both accountable for an action that up to this point the inhabitants of that planet had never known existed?

Consider:

Coincidental Future: In the far future, philosophers determine that it is good not only to make people happy, but also to make happy people. Unfortunately for them, technology at the time is such that it is only possible to generate people randomly. Abiding by a broadly total view of variable population ethics and believing that the expected value of creating people will be positive, they create trillions and trillions of people, 1,000 of whom coincidentally are sufficiently psychologically connected with Hitler. History has been well-documented, and upon learning that the 1,000 are Hitler-clones, society collects a tribunal to blame them.

If the hundreds of clones were the same persons as Hitler, then it is plausible they would still be blameworthy. But these are different people (who cannot possibly be numerically identical to the historical Hitler), who haven’t committed mass atrocities, who were created randomly. These people seem more like scapegoats rather than germane inheritors of past crimes, and any conception of blameworthiness that is consistent with us creating blameworthy people out of thin air seems very

different from the sort of blameworthiness we are ordinarily talking about.¹⁴⁸

While the general psychological connectedness account implausibly allows responsibility to be inherited merely from the psychological similarity resulting from freak coincidence, this does not imply that fission products necessarily lack responsibility for the actions of their originator. And this is because there is something additional about fission products that the general psychological connectedness account does not take into account—they are causally related to their originator. In the next chapter, we turn to the viability of psychological connectedness accounts when we invoke causal constraints on psychological connections.

¹⁴⁸ We can also appreciate how the general psychological connectedness account becomes even more implausible if we try to increase its plausibility by restricting the relevant psychological connections to the values and desires and other psychological states implicated in the original act. This only increases the likelihood that one's becomes responsible for some stranger's actions merely in virtue of coincidentally coming to have similar values.

CHAPTER IX: Causation and the Psychological Account

9.1 Responsibility with Any Cause of Continuity

Khoury and Matheson assert that the diachronic condition is psychological connectedness, which on their view results solely from qualitatively similar psychological elements between two people (the holding of general psychological connections). We saw that this led to the implication that people who coincidentally come to be psychologically similar to others have a chance of inheriting their blameworthiness for past acts. In so far as this is an implausible account of the diachronic condition of moral responsibility, we can anticipate a promising revision to the psychological account by exploring a similar move made in accounts of personal identity. After all, proponents of some psychological criterion of personal identity rarely suggest that there are no constraints on the causes of continuity between the psychological states of two persons. Parfit originally distinguished two versions. According to the wide version, the psychological continuity can arise from any cause, while according to the narrow version, the psychological continuity must come from its normal cause (the continuity of the relevant part of the brain). But what is similar about both versions is that in addition to qualitative similarity, psychological connections don't "count" toward personal identity unless there is some *causal* relationship between qualitatively similar states. I have just argued for the implausibility of a diachronic condition that allows for the transfer of moral responsibility in cases of psychological connectedness that are completely causally unrelated to one another. Can we strengthen the account by adding causal constraints on how two persons come to be deeply psychologically connected?

The most simple refinement of the psychological connectedness account of the diachronic condition would be to say that there has to be at least *some* causal relationship between the two psychological states in order for them to count toward transferred responsibility, but *any* causal relationship will do. Let us call this the:

Wide Psychological Connectedness Account of the Diachronic Condition: If a person A at t_1 is blameworthy to degree d for act X which occurs at t_1 , then an individual B at t_2 is blameworthy for X to degree d if B is maximally psychologically connected to A through wide psychological connections across the

relevant subset of psychological states.

Before criticizing the wide psychological connectedness account, it is worth pausing to note its appeal. The obvious advantage, which needs little explication, is that the account avoids delivering the counterintuitive verdict of responsibility being transferred across only coincidentally connected persons, in at least some cases. The sporadically created thousand Hitler clones in *Coincidental Future* are only accidental inheritors of Hitler's psychological states, and do not deserve blame for Hitler's own actions, nor does the person in *Coincidental Doug*, in virtue of the lack of causal connection between the psychological states that form the relevant connections. A less obvious upshot is that a causal constraint on continuity necessarily links *specific psychological states*, and thus gives the account some resilience to merely accidental causal connections between two persons. It will take a little time to develop this idea while also contextualizing the account for criticisms in the next section, so to start, consider an important question for any proponent of a psychological account of the diachronic condition flagged before and hinted at the end of Khoury and Matheson's paper: "Is diachronic blameworthiness a function of the total psychological connections that hold across time, or does some subset of those connections get special weight?"

We briefly addressed this before, but in more detail, the reason for drawing subsets arises from the observation that, if a one-time criminal is killed and two new people, A and B, are created in his wake, we can imagine that A is entirely psychologically connected to the criminal moments before he died, save for the few psychological connections that constitute his memory of the crime and values leading toward it, while B has those psychological connections, yet is disconnected in many other respects. A is far more psychologically connected with the criminal than B. But B nevertheless seems more intimately associated with the past crime whereas A has none of the malicious evaluative elements that led to the crime in the first place, and hardly seems like an apt candidate for blame. Observations like this invite the view that the diachronic conditions on moral responsibility, in so far

as they depend on psychological connectivity, are not a matter of global connectedness, but rather connectedness across *just* the psychological properties that gave rise to the action in the first place (what I have been calling the relevant subset of psychological states). In other words, what matters for a person's blameworthiness is not his total psychological connectedness to some past person, but rather his psychological connectedness to the mental states of that past person that resulted in the action for which the appropriateness of his present blameworthiness is now in question. This modification nicely dispels any lingering concerns that we are smuggling intuitions about personal identity back into the picture, because the subset of psychological elements implicated in any given action are far too few to constitute the psychological continuity necessary for personal identity if they are connected to some later person.

Now, Khoury's account invokes no causal constraint on how the subset of psychological elements implicated in some past blameworthy action are connected to some future person's psychology, and is thus subject to the previously rehearsed objection that people can (implausibly) become responsible through utterly random coincidences of psychology. As Shoemaker remarks, without a causal constraint "one could find that certain actions are attributable to one solely in virtue of one's utterly coincidental similarity to some stranger who performed them."¹⁴⁹ Indeed, this concern is exacerbated in the subset view because only a small fraction of a person's mental states (those connected to the ones implicated in some past blameworthy action) need to coincidentally align with some distant person's in order to generate blameworthiness, and thus such coincidences will happen more frequently and haphazardly.

However, if we, on the account under consideration, require there to be a causal connection between the two psychological states that comprise that subset of connections related to the act being appraised, we not only dodge this powerful objection to a general psychological connectedness

¹⁴⁹ Shoemaker (2012), 126-127.

account of the diachronic condition, but the way it avoids this is through a dependence relation between the *actual psychological states behind the act* that makes the account avoid seemingly spurious causal connections. For instance, suppose we tweaked our *Coincidental Doug* case where the scientist turns the dial, coincidentally creating a duplicate of the beaten-up body in the original case, only because he smelled a flower that day that put him in an inventive mood, and that flower was planted by Doug. Clearly, the mere causal connection that now exists between Doug and the duplicate could hardly bestow the duplicate with Doug's responsibility now that a totally random connection has been replaced with such a spurious causal one. But presumably Doug would have planted the flower even if the particular psychological states that comprise the connections were different, and so while the connections are causally dependent on Doug, they are not the way they are because his were the way they were.

Before offering general challenge to such accounts, we can appreciate these improvements over the general psychological connectedness account by considering an actual version of a wide psychological connectedness account, Shoemaker's. Shoemaker builds up to some causal dependency by first observing that our intuitions about blaming someone for an action often revolve around the extent to which various actions are expressions of who agents are—that is, they depend crucially on the extent to which an agent in some sense identifies with his actions. But, as Shoemaker argues, there are at least two types of identification we might be referring to, identification with actions, and identification with wills. Identification with an action is usually the simplest to satisfy, and requires some dependence relation between some action and an essential feature of someone's agency, his will. But our wills are usually taken to be “embedded within a much wider volitional web,” and our candidacy for praise and blame is often taken not just to depend on our identification with some action, but also with the will:¹⁵⁰

¹⁵⁰ Shoemaker (2012), 124.

For the self to be sensibly implicated in such appraisals, though, it must now be implicated in the will on which the relevant action depends. Take the self's volitional web to refer to those psychological features that shape, incline, and perhaps even determine its will. The aspects of the self relevant to identification will thus be found in that web. Indeed, the motivating task of many of those dubbed "Real Self View" theorists (see Wolf 1990) has been precisely to identify "the" privileged subset of psychic elements within this volitional web that are truly representative of the agent's "real" self. Various contenders have arisen, including higher-order desires (desires about what one's will is to be) (for example, Frankfurt 1971), evaluative judgments (judgments about actions worth willing) (for example, Watson 1975), and cares (for example, D. Shoemaker 2003). Which one is correct is, again, not my concern here. Instead, I am simply adopting the more neutral stance that it is within one's volitional web that the real self is to be found, and so when that real self is expressed in one's will, one is identified with that will. Consequently, when some action is dependent on that will, one is identified with that action thereby, and it is this relation between an agent's real self and an action that makes the action attributable to the agent.

Now, as we have noted numerous times, even if we judge an agent blameworthy when he commits some action for the reason that he identifies both with it and with the will on which it depends, we usually make such judgments at some later time, and ask the further question of whether that past action is attributable to the person before us. And implicit in that seems to be whether the past action is *still* attributable to the person before us, and hence the natural commitment to personal identity. But Shoemaker believes that fission products who are neither psychologically nor biologically continuous with the original person nevertheless can be responsible for his actions. Given the inability of a personal identity account of the diachronic conditions of moral responsibility to deliver the intuitive verdict to fission cases, Shoemaker suggests that what we really are looking for is a third target of identification: identification with agents, where responsibility can transfer to a current agent in virtue of his identification with the past *agent* who was responsible for the act:¹⁵¹

I have suggested that identification with actions centrally requires a dependence relation running from action to will. I have suggested that identification with wills centrally requires the expression of an agent's self via some shaping or dependence relation between a privileged subset of elements in the agent's volitional web (the real self) and the agent's will. I now want to suggest that identification between a present agent and some past agent centrally requires a similarity relation between the privileged subsets of the current and past agents' volitional webs, a similarity relation that obtains in virtue of a causal dependence relation running from the present agent's real self to the past agent's real self.

In other words, just as a current agent is responsible for an action he commits in so far as it depends

¹⁵¹ Ibid, 125.

on his real self, a current agent is responsible for some past agent's action in so far as his real self is what it is because the past agent's real self was the way it was, where "because" seems to be intended to be understood in terms of some sort of counterfactual dependence.¹⁵² Thus, Shoemaker's account is a wide psychological connectedness account where the relevant subset of psychological states are those implicated in the agent's real self.

9.2 Freak Coincidences and Post-Hoc Retributions, with Causation

While the wide psychological connectedness account does avoid delivering responsibility transfer verdicts to some coincidences, it does not seem resilient enough to this style of objection. We can start by noting that even with respect to any particular psychological connection, it is hard to see how transitioning from no causal relationship between the two states to any relation could somehow imbue the latter with some special properties with respect to such normatively significance concepts as what matters and moral responsibility. Suppose that I used to have a memory of being taken to an amusement park as a child, but I completely forgot about it and no longer have the memory. One day I am hit on the head, which reconfigures (by chance) my brain in such a way as to create a qualitatively identical memory-appearance and a general psychological connection. My new memory-appearance was generated by a process that was causally independent of my previous memory. It seems implausible that this case is relevantly different from one in which I recalled my memory at an earlier age, which my friend remembered for many years, and when he mentions my name because of that memory around some stranger it reminds the stranger of his ex-lover, enrages him and leads him to hit me, spontaneously creating the same memory-appearance. In this case, my new memory-appearance *is* causally dependent on my previous memory and thus is a wide rather than general psychological connection, but there could not plausibly be a difference in counting toward what

¹⁵² I will consider an alternative when evaluating the prospects of what I call the reasons-revision in a later section.

matters or moral responsibility.

Applying this move to moral responsibility does not fare any better. Contrast *Coincidental Future*, where society randomly creates people who turn out to be Hitler-clones, with:

*Coincidental Future**: Hitler's *Mein Kampf* inspires a person far later in time to become a philosopher to better understand ethics in the hopes of shaping a world that created such a monster, and he subsequently publishes the definitive proof of the total view in variable population ethics, as well as empirical research that randomly creating humans will on average bring more positive than negative utility into the world. Work in analytic philosophy profoundly shapes public policy, and society elects to create trillions of random people, 1,000 of which are Hitler-clones.

It is implausible that we would ascribe differential judgments of responsibility or non-responsibility to the Hitler-clones in *Coincidental Future* and *Coincidental Future**. The only difference is that in the later case Hitler's psychological states have some causal relationship to the creation of future Hitler-clones, and this suggests that *mere causal dependency* (of any sort) between psychological connections cannot ground the diachronic conditions of moral responsibility if the psychological connections themselves can't. In fact, trying to tighten the causal constraints on the latter psychological state from any cause that generates it to any cause with some causal dependency on the former psychological state actually seems to bring out why both are implausible views of moral responsibility over time. Consider:

Retributive Future: In the year 3,000, humanity develops the technology to recreate non-historical duplicates of people who lived in the past, who are completely psychologically connected with them. Those in power believe that it was wrong that Hitler never answered for his crimes, and got the easy way out by committing suicide. They do not just want to create one Hitler—their retributive tendencies lead them to create hundreds of Hitlers psychologically identical to the one right before he died in 1945, all of whom they believe are blameworthy for Hitler's atrocities. Moreover, whenever the society learns of another of history's infamous malefactors who didn't seem to get what they deserve, it creates hundreds of clones to blame.

It seems even *more* clear in *Retributive Future* that the Hitler clones are not blameworthy: they are mere scapegoats created for the very purpose of satisfying retributive tendencies. But in contrast to *Coincidental Future*, the original Hitler's psychological states, which determined his actions and subsequently the histories written about him, clearly are in the causal etiology of the Hitler-clones' psychological states, as they are their template. Simply moving from the 'qualitative similarity' view to the 'qualitative similarity with any causal relationship' view cannot vindicate the psychological account

of moral responsibility. As McMahan writes about what matters, “if the *nature* of the causal relation between an earlier state and a later state does not matter, it is hard to see why it should matter whether there is a causal relation at all.”¹⁵³ The same seems true of moral responsibility—the appropriate causal relations that must hold between original and later people’s psychologies are narrower than simply *any* causal relationship.

So, one might reply at this point that the causal constraint is actually more restrictive than simply *any* causal relationship between the paired psychological states that comprise a connection relevant to transferred responsibility. Writing with respect to a randomly created duplicate of, Nozick writes that “though it resembles you, by hypothesis, it does not arise out of you.”¹⁵⁴ Indeed, one of the motivations for imposing a causal constraint, whether for an account of personal identity or moral responsibility, is that the psychological connections don’t seem to count in the relevant sense toward such weighty normative concepts unless the latter *arises*, in some sense, from the former. It is natural to think that you cannot have a latter state arise from or be produced out of a former unless there is a causal connection, suggesting that *if* a latter state arises from a former state, then there is a causal connection between the two. Of course, this by no means implies that *if* a latter state is causally connected to a prior state, then the latter arises from it, and so perhaps we need an even more narrow causal constraint to capture this.

However, it remains far from clear that a principled, non-arbitrary one can actually be specified. Let us start by borrowing from a resource in the literature on personal identity: Kolak and Martin ask us to imagine a supercomputer that randomly manipulates a series of variables and can create MOCs—mathematical object-configurations—that represent real objects and people.¹⁵⁵ A mathematical object-configuration of a person would contain equations specifying all of his neurons

¹⁵³ McMahan (2002), 61.

¹⁵⁴ Nozick (1981), 41.

¹⁵⁵ Kolak and Martin (1987), 343.

and their interconnections as well as the other cells in his body, among other things. The supercomputer has an analyzer that scans a person, constructs an MOC of the person, and then begins randomly making other MOCs, and comparing them with the persons. Perhaps after trillions of random MOCs one will be created that matches the person's MOC, after which it uses the person-corresponding random MOC to create a replica of the person that it simultaneously destroys.

With this technology on tap, let us trace the course of our causal restrictions on psychological continuity and return to the test case that Shoemaker regards as evidence that the personal identity view of transferred (or preserved) responsibility is incorrect: we are supposed to imagine that some criminal assaults another person and then undergoes a botched fission where both products are psychologically (and biologically) discontinuous from the aggressor.¹⁵⁶

Fissioning Faye: suppose Faye assaults a child and then undergoes fission, and because of a botched procedure, both fission products wind up psychologically discontinuous from Faye while still retaining the central aspects of Faye's volitional web that both motivated the original assault and allow the fission products to replay it over and over again with a tingle of excitement. Neither fission product could be identical to Faye on either the Biological or the Psychological Criterion.

In so far as both are responsible (as Shoemaker judges they are), personal identity cannot be doing the relevant work to ground responsibility, and hence Shoemaker defers to his condition that both fission products have a real self that is the way it is in virtue of the aggressor's real self being the way it was. Since Shoemaker clearly intends for there to be some causal dependency, we can rule out the responsibility of cases where the exact same qualitative relation between Faye and two new people obtains, but in which there is no causal relation between her states and the latter's:

Fissioning Faye 6: Suppose Faye assaults a child and is immediately killed. Simultaneous to that, a person-creating supercomputer randomly creates two people who coincidentally are physically and psychologically identical to the fission products in *Fissioning Faye*, and who thus replay the events of the assault with pleasure and share a structurally similar real self to the one Faye expressed in her assault.

In the first revision we considered, the *Wide Psychological Connectedness Account*, Faye's fission-equivalent products would not retain her responsibility since there is no cause of their continuity that relates to

¹⁵⁶ Shoemaker (2012), 125.

Faye or Faye's psychological states. Of course, that account fails to deliver the intuitive non-responsible verdict if we suddenly rig some fringe causal connection between them, such as in:

Fissioning Faye 5: Suppose Faye assaults a child and is immediately killed. Simultaneous to that, a person-creating supercomputer randomly creates two people who coincidentally are physically and psychologically identical to the fission products in *Fissioning Faye*, and who thus replay the events of the assault with pleasure and share a structurally similar real self to the one Faye expressed in her assault. The supercomputer does this because a glitch in the program makes it act in such a way in the presence of a particular flower, and that flower was planted by someone in disgust reacting to Faye's real self.

I have already argued that we should treat cases like *Fissioning Faye 5* and *Fissioning Faye 6* symmetrically—the newly created people in both cases should either all be non-responsible or all be responsible, as there is no relevant difference to responsibility between no causal relationship and such spurious causal relationships. Let us contrast this with a case like:

Fissioning Faye 2: Suppose Faye assaults a child and then intentionally undergoes fission, and because of a botched procedure (that perfectly copies the psychological states but not the physical material), the actual underlying physical brain tissue is replaced and the products wind up psychologically discontinuous from Faye while still retaining the central aspects of Faye's volitional web that both motivated the original assault and allow the products to replay it over and over again with a tinge of excitement.

A proponent of the wide psychological connectedness account is likely committed to agreeing that the products in *Fissioning Faye 2* are responsible for Faye's assault. This is because the case is designed to be like *Fissioning Faye* in all aspects except the actual physical continuity of Faye's brain tissue to the new persons. Given the fact that the proponent of the wide account won't think that physical continuity is necessary (otherwise they would endorse the normal psychological connectedness account of the diachronic condition, which I discuss in the next section), and given that a normal psychological connectedness account is even more causally restrictive, it seems like you cannot get any more causally restrictive on the causes of psychological continuity before the account actually becomes a normal connectedness account. But if this is the test case for which any wide psychological connectedness proponent must concede that responsibility is transferred, then it is unlikely that any restrictions to the wide account actually can make it more plausible than the general account. That is because we should judge *Fissioning Faye 2* symmetrically with:

Fissioning Faye 3: The same as in *Fissioning Faye 2*, but Faye accidentally rather than intentionally triggers her own botched pseudo-fission.

The mere fact that Faye has an intention to undergo a biologically discontinuous version of fission could not plausibly ground a differential responsibility verdict delivered to the emergent products in *Fissioning Faye 3* compared to *Fissioning Faye 2*. But then, consider:

Fissioning Faye 4: The same as in *Fissioning Faye 3*, but when Faye accidentally hits the button, the supercomputer it generates billions of MOCs randomly, and when one randomly generated MOC happens to match Faye's MOC it creates two of them.

To clarify, the difference is that in the previous cases, when Faye accidentally (or intentionally) hits the button, the supercomputer scans her and creates an MOC from which it instantiates the products by physically creating them. Could the fact that an MOC is created on the basis of a person's scan and then instantiated versus created randomly and instantiated on the basis of a person's scan actually ground a difference in whether responsibility is sustained? Writing about this as it relates to responsibility, Kolak and Martin find the prospect of a difference between the two to be absurd¹⁵⁷, and indeed it is hard to see how such a slight nuance in the causal order could have such a profound impact on the blameworthiness of newly created people. But if that is true, then the wide psychological connectedness account is in hot water, because it is hard to see why we should asymmetrically judge *Fissioning Faye 4* and *Fissioning Faye 5*. The difference between them, of course, is where in the process of creating the new person Faye's original states impute their causal role. In all of these cases, there are essentially two steps to the creation of the Faye duplicate: (1) the electronic creation of an identical MOC, and (2) the physical instantiation where the electronic MOC is turned into a living, breathing being. In *Fissioning Faye 4* the causal connection happens during (1) since the MOC that is generated is generated based on the scan of Faye, whereas in *Fissioning Faye 5* the causal connection happens during (2) since the MOC is generated randomly but the physical instantiation of the MOC occurs

¹⁵⁷ Kolak and Martin (1987) 342.

because of the flower's causal contribution. Again it seems arbitrary that which step is random and which step receives causal input from Faye's original states should make such a large difference to the responsibility of the newly created persons, yet if this is true, then we have arguments that symmetrically link all the cases from *Fissioning Faye 2* (the tightest causal constrain before adopting a normal cause account) to *Fissioning Faye 6* (where there is no causal connection whatsoever); the wide psychological connectedness account of the diachronic condition is no more plausible than the general account

That fission products do not sustain the responsibility of their antecedent actor is by no means fatal to Shoemaker's account. It perhaps makes the account less motivated if the inability for the personal identity account of the diachronic conditions of moral responsibility was the main reason for exploring an alternative—such fission products are not responsible after all. To review, all I have argued in this section is that anyone who believes that responsibility transfers to fission products should either abandon any causal dependency constraint relating current and past real selves, or explain the importance of physical continuity. If the constraint is given up, then Shoemaker must endorse the view which he explicitly rejects—that causally unrelated doppelgangers could inherit responsibility merely in virtue of coincidental psychological similarity. This at least makes it structurally clear what recourse is left for the proponent of a psychological connectedness view who rejects the general version: either try to articulate why the physical continuity between fission products makes a difference in transferring responsibility over the pseudo-fission cases just discussed, or articulate some other causal constraint that may not even apply to cases of fission. We turn to both of these in the next two sections, respectively.

9.3 Normal/Narrow Causes

I have argued against a psychological connectedness conception of the diachronic condition of moral responsibility if it is cashed out without a causal constraint, and if it cashed out with a wide cause

constraint. At this point it is natural to think that the causal constraints examined so far are still too wide, and that, instead, there is a narrow or “normal” causal requirement on psychological connectedness; responsibility is only preserved when psychological connectedness is preserved through the usual way that connectedness is realized—i.e. continuity of the relevant part of the brain.

So let us evaluate a normal cause view of the diachronic condition of moral responsibility, such that you can only be responsible for some past action if there is the normal cause of continuity of the psychological elements behind the action—continuity of the relevant parts of the same brain. This appears to be the view that is consistent with McMahan’s in his discussion of what matters. In so far as moral responsibility is part of or supervenes on what matters¹⁵⁸, McMahan agrees with the earlier arguments that there must be some normal-cause causal requirement on the psychological connections, since he believes there is no relevant difference with respect to the preservation of what matters between intrinsically connected psychological states that bear no causal relation to one another and ones which simply bear some causal relationship. Since he does not believe that what matters is preserved in the former, nor does it in the latter, and thus the causal constraints must be more restrictive than merely positing *any* cause:¹⁵⁹

[M]any of us doubt that the mere replication of a mental state establishes a psychological connection or that pseudo-connections established via replication would constitute a basis for egoistic concern. If we are to accept that psychological connectedness is part of what matters, we might therefore insist on a concept of a psychological connection according to which an earlier and a later psychological state form a psychological connection only if the two are qualitatively similar and the later state is causally dependent on the earlier one in the right way. And, we might further insist, there is the right sort of causal dependency only if there is continuity between the physical realization of the earlier state and the physical realization of the later state—that is, only if the states are constituted or generated by the same region of the same brain.

With this, McMahan invokes a narrower conception of a psychological connection—a *real* psychological connection—in which the existence of a psychological connection (that could form the

¹⁵⁸ In correspondence, McMahan has expressed that he accepts the claim that moral responsibility supervenes on what matters.

¹⁵⁹ McMahan (2002), 63.

basis of the preservation of what matters) depends not only on qualitative similarity between two states and not only on some causal requirement but for that causal relation to be the normal cause of the continued existence and function of enough of the same brain.¹⁶⁰ This is why, while McMahan believes there is no preservation of what matters in cases of replication, there are in fission cases and of course vanilla cases of sustained existence over time, where the psychological continuity is sustained by the continued existence of the same brain tissue that the original psychological states were realized in. We can call this the:

Normal Psychological Connectedness Account of the Diachronic Condition: If a person A at t_1 is blameworthy to degree d for act X which occurs at t_1 , then an individual B at t_2 is blameworthy for X to degree d if B is maximally psychologically connected to A through real (normal) psychological connections across the relevant subset of psychological states.

We now no longer need to worry about the bizarre cases of people being spontaneously duplicated by supercomputers, because such causal connections are hardly normal. Nor do we need to worry about cases where people are only spuriously causally connected to others. Unfortunately, I think this runs into a difficulty that we can pinpoint by considering cases where agents do have normal cause-preserved psychological connections yet do not appear to be responsible. Consider a case like:

Disaggregated Doug: Doug freely kills another person (he satisfies the synchronic conditions of responsibility with respect to that act of killing). The values implicated in this action, as instantiated by the functional role of some collection of his neurons, persist, but neuroscientists replace the remainder of his brain with qualitatively identical neural substances such that the new person is indistinguishable from Doug.

Neuroscientists can completely finagle other parts of Doug's brain without disrupting the normal psychological connectedness and continuity of the subset of psychological elements tied to his murder, and I take it to be a relatively noncontroversial empirical assumption that a collection of his neurons whose functional role comprise the values implicated in the murder does not comprise the majority of Doug's brain, or anything close to it (even 10% seems like a generous assumption).

¹⁶⁰ Ibid.

Since this is nowhere near a plausible threshold of a proponent of the psychological criterion of *personal identity*, on the psychological view of personal identity, there can be connectedness and continuity of the psychological elements related to Doug's murder without strong connectedness, and hence without sustained personal identity (and, to make a concession to the biological criterion, we can imagine that the brain is transplanted to a different body such that personal identity is not sustained on a biological account either).

At first, this seems exactly the implication expected. Doug's identity isn't preserved but responsibility is transferred to his successor, and so we have responsibility without personal identity. Moreover, it is on the basis of connectedness of the relevant subset of his psychological elements that the successor is responsible for Doug's past actions. But when we look at it a different way, really what has happened is that we have created an instant agent, part of whose brain is Doug's brain—the very part associated with his murder. And so if we zoom in on that, we will find that responsibility is preserved in:

Disaggregated Doug 2: Neuroscientists are creating an instant agent. They have 90% of the brain ready, and take the 10% of Doug's brain that, in *Disaggregated Doug*, was causally implicated in his murder. Call this instant agent Doug*.

While Doug* is undeniably a different person from Doug, and while Doug* is psychologically disconnected from Doug, Doug* *is* psychologically connected, by the normal cause, with Doug *with respect to a subset of his psychological elements*, and these are the very ones that were associated with the murder. So, on the normal cause variation of a psychological connectedness view of moral responsibility over time, Doug* is responsible for Doug's murder. But contrast this with:

Disaggregated Doug 3: Neuroscientists are creating an instant agent. They have 90% of the brain created, and take the 10% of brain material that is qualitatively identical to the 10% of Doug's brain in *Disaggregated Doug 2*. Call this instant agent Doug**.

Are we to believe that Doug**, because he received a transplant of organic matter that was qualitatively but not numerically identical to Doug's brain, is *not* responsible while Doug* is? It is hard to see why

brain material that plays the exact same functional role could differentially affect Doug* vs. Doug**'s responsibility depending on whether it is made of the same neurons as Doug's.

But while that may be a way of describing the cases that makes it sound odd to think that the difference between the two could be important, perhaps upon further consideration the difference in fact is not so arbitrary. Indeed, McMahan writes:¹⁶¹

We may concede that there would be psychological connections between you and me if parts of my brain containing memories and other elements of my mental life were excised and grafted onto your brain. And we might concede, further, that if enough such matter were going to be transplanted from my brain to yours, I would have grounds, although relatively weak ones, for being concerned in an egoistic way about what would later happen to you.

After all, the instant agent in *Disaggregated Doug 2* shares real psychological connections with Doug (in virtue of there being continuity of Doug's brain) whereas the instant agent in *Disaggregated Doug 3* does not. But while invoking real psychological connections does appear to be a principled way to distinguish cases of fission and ordinary persisting existence (where what matters is putatively preserved) but not teletransportation (where it is not), it seems implausible that this could serve as an important difference with respect to moral responsibility. To see why, let us look more closely at various concepts of continuity that McMahan uses:¹⁶²

Physical Continuity: Continued existence of the same matter or gradual replacement of it over time.

Functional Continuity: Continuity of the brain's psychological capacities.

Structural Continuity: Continuity of the organization that underlies particular psychological contents.

There is a question of which of these types of continuity can be determined to obtain solely by comparing the intrinsic features of two states without knowing anything about their causal relationship. On this question, McMahan is unequivocal. Clearly, we cannot say that there has been physical continuity between two brain states at different points in time that are qualitatively identical, since we don't know whether the same brain stuff that comprised the former now comprises that

¹⁶¹ Ibid., 63.

¹⁶² Ibid., 68-69.

latter. But the continuity of the psychological capacities and psychological contents solely depends on the qualitative features of two states, which is why McMahan says that “there is functional and organizational continuity of the brain when a person is replicated (as in Teletransportation), even though there is no physical continuity.”¹⁶³ However, it appears that a significant type of continuity is left out here, one that emerges in the case of:

Unscrambling Transportation: Unlike in the case of Teletransportation, where one is instantaneously disintegrated and replicated out of new matter to be a cell-for-cell duplicate, upon entering a scanning booth and pressing the button one is instantaneously disintegrated but rather than the information being transmitted by the speed of light to the replicating booth elsewhere, the actual atoms are transmitted near light speed and the Replicator creates a qualitatively identical construction of one’s body with the original atoms (including the original brain tissue).

While the case of Teletransportation has, as McMahan flags, both functional and organizational continuity (in virtue of the replica being a cell-by-cell duplicate) but not physical continuity, *Unscrambling Transportation* has physical continuity in addition to functional and organizational continuity. However, if one is disinclined, as McMahan is, to believe that moral responsibility is preserved in the case of teletransportation, it is natural to think that he could also plausibly have doubts about whether it is preserved in unscrambling transportation cases, while the account in question implies that responsibility is preserved.¹⁶⁴ While the specific atoms that comprise someone’s brain certainly are relevantly different than qualitatively identical ones *when they comprise that brain*, after they are disaggregated to individual atoms it is hard to believe that there is something deep in the metaphysics of moral responsibility that maintains a ledger about whether particular discrete atoms were once part of some brain; if you remove atoms from their functional role they quickly become organic marbles that are interchangeable with one another.

The distinct type of continuity being hinted at here seems to be what might be called *functioning*

¹⁶³ Ibid.

¹⁶⁴ McMahan has confirmed in correspondence that he accepts the view that the new person is responsible in *Unscrambling Transportation*.

continuity, which is to (what McMahan dubs) functional continuity as real psychological continuity is to non-causal broad psychological continuity. That is, a latter psychological state can be broadly continuous with a prior one just by being qualitatively similar, but only in the case of real psychological continuity is the stuff that comprises the two (qualitatively identical) the *same stuff*. Similarly, while a latter psychological state can have functional continuity with a prior one just by being qualitatively similar (so that the two both share a capacity for consciousness), only in the case of *functioning* continuity are the functions that comprise the two qualitatively identical functions the same. So while there may be functional continuity between my replica and me in the teletransporter case in virtue of us having relevantly similar capacities for consciousness, his capacity is not my capacity and so we lack functioning continuity. (Likewise, the atoms that undergird his psychological states are not my atoms, and so we lack physical continuity). Importantly, this lack of functioning continuity is not overcome when we move to a case of unscrambled transportation. Although physical continuity is undoubtedly established since the atoms of the replica are my atoms, plausibly the particular capacity for consciousness that was destroyed when my atoms were unscrambled does not come back into existence when the replica is built up again, but replaced by a very similar capacity.

Now my goal here is not to wade into the metaphysics of capacities or to offer arguments directly on the correct account of what matters—such matters far exceed the scope of this thesis. It is merely to illustrate considerations about continuity in the context of cases usually implicated in the dialectic on what matters which I think have direct relevance to diachronic considerations about moral responsibility. While arguments I make about transferred responsibility between two persons can likely be made regarding what matters (though perhaps with differential force), my concerns here are far more tactical. And while I believe people could reasonably differ on whether the sort of brain graft that McMahan describes (which preserves physical and functional, but not functioning, continuity) might give one weak grounds for being concerned in an egoistic way about what happens to the

recipient of the graft, it seems implausible in the context of moral responsibility. Suppose that we are in a tribunal evaluating the moral responsibility of a man who has consensually received a brain graft of the physical realization of some psychological state underlying a prior free and blameworthy act. In this world, society has scanned the brains of criminals who for one reason or another died after freely committing some crime, and has not only interpolated what the exact configurations of their brains were when they freely acted but has also archived the very brain tissue that physically realized the states they were in when they acted. One can not only see a giant repository of this archived brain tissue, but if one is so inclined receive a graft of the original neurons, in the right configuration, that realized some blameworthy psychological state (i.e. the one that undergirded the blameworthy action). As the court learns that this man has willingly taken up this graft, the tribunal determines that according to the normal psychological connectedness account of the diachronic condition of moral responsibility, the man is guilty. But wait—the defense attorney now provides evidence that the night before the graft, some sneaky teenagers broke into the facility. Wanting the original brain tissue of a criminal they admire, they managed to steal the original archive of the graft the defendant received, and replaced it with qualitatively identical brain tissue, so nobody can tell that any theft has happened just on the basis of examining the tissue itself. The tribunal orders a thorough investigation of the alleged theft; after all, if it did happen, the defendant may not be blameworthy after all...

I take it to be implausible that we cannot make a decision about the man's blameworthiness without knowing whether some pesky teenagers replaced a few atoms with identical ones. While this is really just a more elaborate version of the original difference between *Dissaggregated Doug 2 and 3*, now we can say more: the *reason* it seems like physical continuity is important in transfer of responsibility is because you cannot have transferred responsibility without functioning continuity, and you cannot have functioning continuity without physical continuity. So physical continuity may be coextensive with the cases in which moral responsibility is transferred but it is not the criterion for

determining that it has.

Why is this problematic for a proponent of some normal cause variation of the psychological connectedness account of moral responsibility, rather than just a nudge toward a certain sub-variation of it? Whether it is the former or the latter depends on how we adjudicate between two competing views. According to one, the relations constitutive of personal identity are necessary and sufficient for the functioning continuity that I have argued is needed for transfer of moral responsibility. That is, for one to have functioning continuity with some past person is to satisfy the minimal conditions for personal identity (because, for instance, the individual must have, on McMahan's criterion, "enough of one's brain to be able to support consciousness"¹⁶⁵). According to the other view, the relations constitutive of personal identity are sufficient but not necessary for functioning continuity; while cases of ordinary persistence of personal identity do indeed sustain functioning continuity, such functioning continuity can exist between two persons even in the absence of shared identity. The adjudication of these two views obviously cannot be adequately done without first explicating the correct account of personal identity and whether it is either determined by or coextensive with functioning continuity. But carrying out this task is not needed to accomplish my goal in this chapter; in trying to build the strongest case for a psychological connectedness account of diachronic responsibility, we tried to build out clear cases where intuitively responsibility is transferred or preserved even when identity clearly isn't. Failing to do so when psychological connectedness had no causal constraints, we began adding causal conditions, down to the normal cause, with no luck. The one variation that does not, as I've argued, face strong complications—one involving functioning continuity—no longer clearly proffers cases where responsibility is preserved but identity isn't, because functioning continuity might very well entail personal identity. *Prima facie*, functioning continuity demarcates continuity of the same consciousness and plausibly cannot exist in the absence of sustained personal identity, and so if

¹⁶⁵ Ibid., 85.

functioning continuity is the feature of some of the cases described that allows for transfer of responsibility (or sustained responsibility), the dialectical burden is on the proponent of the account in question to argue how you can have functioning continuity without persistence of personal identity.

9.4 The Reasons-Revision

If the arguments in the last few sections have been correct, then one of the lessons is that increasing causal *intimacy* alone is not as compelling a move when used for the diachronic condition of moral responsibility as it is when used for the psychological criterion of personal identity. Perhaps if there is some merit to a causal constraint, it won't be by going from general to wide to normal causes, but by specifying some other *type* of causation that does not cut along this wide to narrow spectrum.

When he sketches out the rough contours of a causal dependency constraint at the end of the paper, Shoemaker offers a suggestion that contains the outlines of a promising causal dependency constraint available for subset views:¹⁶⁶

One might think, though, that this [causal] component could then be met only in virtue of the relevant psychological states being manifested from their persistence in the same brain (or relevant part of the brain), but this would be too narrow a conception. I think it is important to allow for the possibility of my identification with other agents (even strangers), those who do not share my brain but who nevertheless have had a significant causal impact on the formation of my real self, through my interaction with them or their ideas. To the extent that their volitional selves have become mine by this causal route, then, so too have their actions.

When Shoemaker notes “significant causal impact,” he goes further than this to note that the impact arises “through” an “interaction.” Though he never specifies what he means by this, it is suggestive of a causal constraint sensitive not just to the magnitude of impact that some previous real self has on an agent's current real self, but the *interactive way that has come about*. What comes to mind is not some brute mechanism by which a past real self can causally impose itself on a current real self in any way, but rather one in which a current person's psychology comes to be relevantly like some past self *through* an understanding and appreciation of why that past self is the way it is—that is, it is sensitive to *reasons*

¹⁶⁶ Shoemaker (2012), 127.

for becoming like that past self. Let us call this revision to the psychological connectedness account the reasons-revision: Shoemaker specified identification with some past agent as obtaining in virtue of a similarity between a current agent's real self and a past agent's real self, where the current real self is the way it is *because* the past self was the way it was; on the reasons-revision, identification with some past agent obtains in virtue of that very similarity, but the current real self is the way it is *because of the reasons for being* the way some past self was.

Reasons-Revision of the Psychological Connectedness Account of the Diachronic Condition: If a person A at t_1 is blameworthy to degree d for act X which occurs at t_1 , then an individual B at t_2 is blameworthy for X to degree d if B is maximally psychologically connected to A through reasons-bearing psychological connections (where a psychological connection is reasons-bearing if the latter psychological states are caused to be the way they are for reasons related to the former states being the way they are).

It is clear on this view that neither psychological nor biological continuity is required; as Shoemaker would like, a substantive interaction with another person or his ideas may be sufficient to come to identify with him such that one inherits responsibility for his actions. Moreover, this allows us to draw a wedge between such responsibility-grounding interactions and the need to uphold the responsibility of fission products, since the fission products may be the way they are because (and even primarily because) of some past self, but they have by no means come to be that way by tracking the reasons for being that way. The same would apply to exclude the coincidental matches and bizarre doppelgangers that Shoemaker wants to exclude, whose real selves do have a similarity but not by virtue of engaging with some past real selves.

Appealing to reasons almost by definition allows us to draw a non-arbitrary line when differentiating various causal relationships between past and current real selves, since we allow transferred responsibility only when an agent's real self develops through appreciation of what counts in favor of being like the past real self with which it interacts. However, despite its promise, I believe this account also has severe complications. Imagine what I take to be the sort of exemplar case Shoemaker has in mind when he imagines us inheriting responsibility by interacting with someone and

their ideas:

Hitler's Protégé: Hitler is responsible and blameworthy for a specific incident where he personally orders the deaths of hundreds of Jews standing before him. An ordinary man named Clyde becomes enamored by Hitler and comes to shadow him everywhere, taking deep inspiration by interacting with Hitler and coming to believe his values. Clyde's real self comes to be relevantly similar to Hitler's, which was expressed in the specific incident.

Shoemaker asserts that "To the extent that their volitional selves have become mine by this causal route [of interacting with them or their ideas], then, so too have their actions." If there was ever an instance when such an effect occurred, it occurs with Clyde, and hence it seems that Shoemaker should agree that he identifies with Hitler's real self in such a way that he is responsible for the ordering of the execution of hundreds of Jews. I happen to not share this intuition, but appealing to a tug of competing intuitions hardly pushes the dialectic forward. On the other hand, it does seem like the reason for upholding Clyde's responsibility in *Hitler's Protégé* should apply to:

Hitler's Protégé 2: Hitler is responsible and blameworthy for a specific incident where he personally orders the deaths of hundreds of Jews standing before him. An ordinary man named Clyde becomes enamored by Hitler simply by reading *Mein Kampf*, which results in his real self being identical to the one in *Hitler's Protégé*.

If the interaction with Hitler and his ideas indeed can make Clyde's real self relevantly similar such that he is responsible for the mass execution, it should hardly make a difference whether this same effect is achieved through interacting with Hitler in person or through his book. Nor does mere clock time seem like a relevant factor in the responsibility-grounding effects of such interaction:

Hitler's Protégé 3: Hitler is responsible and blameworthy for a specific incident where he personally orders the deaths of hundreds of Jews standing before him. Five hundred years later, an ordinary man named Clyde becomes enamored by Hitler simply by reading *Mein Kampf*, which results in his real self being identical to the one in *Hitler's Protégé*.

I find it incredible that Clyde could become responsible for an act that was committed before he was even born, but Shoemaker very clearly intends to allow for this possibility. However, an even stranger result emerges when we note that it does not seem like it matters whether *Mein Kampf* was written before or after the specific mass execution. As a historical matter, it must have been written before any executions Hitler ordered, but *as long as his real self that was expressed in the memoir remains the same*, its

causal effects on Clyde seems unchanged. But then we have three time points of interest:

- (1) When Hitler orders the mass execution expressing his real self.
- (2) When Hitler writes *Mein Kampf* expressing the same real self that orders the execution.
- (3) When Clyde reads *Mein Kampf* and comes to have, for reasons, a real self that is identical to Hitler's in (1) and (2).

Clearly, for the memoir to have such an effect on Clyde, (3) must come after (2). But we have noted that it does not matter, for (3) to causally depend on (2) in the requisite sort of way, whether (1) or (2) comes first. Therefore, Clyde could actually come to be responsible in:

Hitler's Protégé 4: Hitler writes *Mein Kampf*, expressing the same real self that will later identify with the action and will of ordering a mass execution. A day before Hitler orders the execution, Clyde reads the memoir and comes to have an identical real self.

It does matter for his responsibility, whether Clyde reads the memoir a day before or a day after the execution, but this means that we could come to have real selves similar to other agents, and then *become responsible* for actions that they have not done yet, but will do in the future. Indeed, on this view, as soon as someone's real self has influenced my real self in the specified way, I suddenly become responsible for not only every action that previously stemmed from that real self, but I will eventually become responsible for all future actions stemming from that self as well.

If this is not a decisive enough complication with the account or one is unpersuaded of its soundness, we can home in on one other source of arbitrariness by considering:

Hitler's Protégé 5: Hitler writes *Mein Kampf*, expressing the same real self that orders a mass execution. However, Hitler has a causally unrelated historical duplicate who also writes a word-for-word identical memoir, but who subsequently underwent far different life experiences and lived an innocuous and relatively blame-free life. Five hundred years later, Clyde reaches for a copy of *Mein Kampf* on a bookshelf that contains one written by Hitler and the other written by his duplicate. He grabs the one by Hitler, and develops, through interacting with Hitler's ideas, a sufficiently similar real self.

Hitler's Protégé 5 could very well be the same story (with a few extraneous details) as *Hitler's Protégé 2*, and so we ought to, according to the reasons-revision account, judge him responsible for Hitler's atrocity. However, if Clyde had instead grabbed the copy written by the crime-free duplicate, or one generated by monkeys randomly jabbing at keys on a computer, he would not be responsible for Hitler's crime; Clyde came to have a real self that did respond to reasons, but such reasons were

causally unrelated to *Hitler's* being the way he was. If Clyde was not causally influenced by the *actual real self* that was expressed in some blameworthy act, he is not responsible, yet the belief that which book Clyde grabs could so greatly affect his responsibility is seems implausible. Even if the reader is unpersuaded by this point, it is worth pointing out that in requiring the cause of continuity to be reasons-bearing, the proponent of the reasons-revision also faces difficulty in articulating how responsibility is sustained in ordinary cases of persisting identity. After all, it is uncontroversial that if we are morally responsible for our actions at all, we are morally responsible moments after committing them. But the normal cause of continuity is not *prima facie* a reasons-bearing cause, and so it is on the burden of the proponent of the account in question to say why it is, or else why responsibility is sustained in such ordinary cases.¹⁶⁷

9.5 A Universally Troublesome Case for Psychological Connectedness Views

While I believe that the arguments in this chapter provide strong evidence against the seemingly most attractive variations of psychological connectedness accounts of the diachronic condition, there is one more difficulty that, rather than applying to specific versions (i.e. the normal cause variation), applies to all of them, perhaps even providing a positive argument for the personal identity account. It is an expansion of an argument intimated in 7.4, and begins with a former case:

Drunk Doug: Doug freely gets drunk at a bar and drives home. He lacks even basic coordination, momentarily blacking out and hitting a pedestrian on the side-walk before waking up to an airbag in his face. He is essentially blacked out at the moment he hit the unfortunate chap, and *at that time* does not satisfy the requisite non-historical components of the synchronic conditions of moral responsibility.

I bring up *Drunk Doug* as a reminder of the basic move used to preserve the responsibility of agents for actions for which the synchronic conditions of moral responsibility are not satisfied at the time they are performed. The move is that Doug was responsible for getting drunk, and could have or should have foreseen this possible outcome, and so his responsibility for hitting the pedestrian derives

¹⁶⁷ Thanks to Alison Hills for this point.

from his previously free act. It seems that a structurally similar move can be made to accommodate cases such as:

Responsibility-Dodging Doug 1: Doug freely kills another person (he satisfies the synchronic conditions of responsibility with respect to that act of killing). He learns about the psychological connectedness account, one that has become readily accepted by most people, and believes he has made a realization: he is off the hook as soon as he gets rid of the subset of evaluative elements that were causally implicated in the relevant way to his killing. He accesses a machine that can eradicate and replace discrete evaluative elements, and removes the values in the causal etiology of his killing, the memory, etc...

The similar move is that just as responsibility for what *is* done can derive from some past free action, so too can responsibility for what *was* done. In so far as the proponent of the psychological connectedness view shares my intuition that Doug cannot evade his responsibility by deliberately engineering away certain elements in his evaluative scheme, he may reply that Doug *is* responsible for that act of removing certain values, and his responsibility for what he did *derives* from that prior free action.

I am willing to grant that this is an acceptable response by the psychological connectedness theorist for Doug's persisting responsibility for his murder. In general, it seems like a very powerful tool for any Ulysses-esque cases where a person deliberately alters his own values in the service of some goal. But this derivatives tool is inadequate for cases such as:

Responsibility-Dodging Doug 2: Doug freely kills another person (he satisfies the synchronic conditions of responsibility with respect to that act of killing). He learns about the psychological connectedness account, one that has become readily accepted by most people, and believes he has made a realization: he is off the hook as soon as he gets rid of all of the evaluative elements that connect him to the crime, *and* that connect him to his wanting to disconnect himself from the crime. He accesses a machine that can eradicate and replace discrete evaluative elements, and has it do two things. First, it removes the values in the causal etiology of his killing, the memory, etc...Second, it removes the values in the causal etiology of his deciding to use the machine to do these things, and the memory of that, etc...

In this case, Doug not only removes the values related to his killing, but also those related to that very removal. It is not plausible that by additionally removing the values that led him to do the very action of removing his values that led him to kill, Doug is any less responsible for the murder than he was in *Responsibility-Dodging Doug 1*—and so Doug is not able to evade his responsibility. Now, the

psychological connectedness theorist does *not* seem to be able to account for why Doug is still responsible. He cannot derive responsibility for killing (to which he has no psychological connectedness) by way of his decision to eradicate the psychological connections he had with the killing, because he has also eradicated any psychological connections he has with the decision itself (to which he now has no psychological connectedness). A proponent of personal identity as the diachronic condition of moral responsibility would be able to use Doug's persisting numerical identity over time to glue together his responsibility through time, but this option is unavailable according to any psychological connectedness account which disputes the necessity of identity. In so far as we think Doug is responsible, it is hard to imagine a feature other than identity grounding it.

PART III: Intersecting the Diachronic and Synchronic Conditions
CHAPTER X: The General Psychological Account and Synchronic Conditions
10.1 Permutations

The first two parts of this thesis stand relatively self-contained, and while intersections have occasionally popped up, they have remained little explicated and largely unexamined. The first part explored the question of what makes it the case that we are responsible for what we do, and the second part explored the question of what makes it the case that we are responsible for what's been done. The question of responsibility at the time of action largely broke down into two camps: the historicists and the non-historicists, the former of which most of our time was spent exploring. The question of responsibility for past action largely broke down into two camps as well: the personal identity account and the psychological connectedness account, the latter of which most of our time was spent exploring.

For each of these sets, we have peeked inside each camp as well as contrasted one camp against the other. But where on the responsibility battleground do these two sets of camps lie in relation to one another? Can one mix-and-match and coherently spin the dial for any position about the synchronic conditions of moral responsibility and any position about the diachronic conditions? Or are certain camps across each set incompatible with one another—does occupation in one camp preclude occupation in another?

In this final part of the thesis, I turn away from the absolute and relative plausibility of the various synchronic and diachronic conditions of moral responsibility, and toward the plausibility of various permutations of these conditions. I begin by exploring each conceptually possible permutation among the psychological connectedness views: whether (in this chapter) we can conjoin the general psychological connectedness account of the diachronic condition with either non-historicism or historicism, and whether (in Chapter XI) we can do the same with the reasons-revision of the psychological connectedness account. I argue that on top of any complications with each account

separately, serious interactive complications arise when we mix any psychological connectedness account with any synchronic account. Next, in Chapter XII, I turn to the personal identity account of the diachronic condition, and argue that it also faces interactive problems when paired with non-historicism. Only the personal identity diachronic account paired with historicism seems to lack obvious conjunctive problems.

10.2 The Psychological Account and Manipulated Agents

Let us home in on two of the camps we have encountered so far. One, a candidate account for the diachronic conditions of moral responsibility, states:

The General Psychological Connectedness Account: If a person A at t_1 is blameworthy to degree d for act X which occurs at t_1 , then an individual B at t_2 is blameworthy for X to degree d if B is maximally psychologically connected to A through general psychological connections across the relevant subset of psychological states.

The other, a candidate account for the synchronic conditions of moral responsibility, states:

Historicism: An agent is morally responsible for some action if and only if he meets certain non-historical and historical criteria.

In a trivial sense, there is nothing logically inconsistent between these two positions, and so it seems perfectly kosher for someone to believe both. But, as I will now argue, some serious problems lurk below the surface. To evaluate a combination of the general psychological connectedness view and historicism, let us return to a case we encountered previously:

Brainwashed Beth: When Beth crawled into bed last night she was an exceptionally sweet person, as she always had been. While Beth slept, a team of psychologists that had discovered the system of values that make Chuck [a vicious killer] tick implanted those values in Beth after erasing hers. Beth devises a plan for killing another person, and she executes it—and him—that afternoon.

It is for cases exactly like *Brainwashed Beth* that most historicists go historical in the first place; even if she happens to satisfy the requisite non-historical criteria for responsibility at the time of murder, Beth does not seem responsible, and historicism is needed to vindicate that intuition. But now we can supplement the thought experiment with some color pertaining to Beth's responsibility for past acts:

Brainwashed Beth (continued): Not only did the psychologists implant Chuck's values into Beth, but they also implanted (false) memories of Chuck's latest killing, which was done wholly in virtue of the very values that were planted into Beth. This is done in such a way that Beth is (close to) maximally psychologically connected at t_2 to Chuck when he murdered his most recent victim at t_1 .

According to the most general version of the psychological connectedness account, when Beth awakens at t_2 and "remembers" freely killing Chuck's victim, Beth is responsible for that murder. This is a severe complication for both positive and negative varieties of historicism because while historicists might (if they are disposed to negative historicism) ascribe freedom to the incipient acts of instant agents, neither positive nor negative historicists will want to blame the incipient acts of manipulated agents. And, as I will now argue, it is deeply implausible that brainwashed Beth could be responsible for Chuck's latest killing but not the one she immediately goes on to do. That is, one might plausibly hold (if he accepts the general psychological connectedness account) that Beth is responsible for Chuck's latest killing. And one might plausibly hold (if he accepts historicism) that she is not responsible for her incipient actions post-manipulation. But one cannot plausibly believe the conjunction of the two because if Beth is responsible for Chuck's latest killing, she must be responsible for her own.

The claim being made is that our answer to whether Beth is responsible for Chuck's actions has a connection to our answer to whether Beth is responsible for any future actions she commits. To see why one might be drawn to this claim, consider the following principle:

Principle of Comparative Autonomy 1 (PCA1): If a person is responsible for *someone else's* action in virtue of satisfying certain psychological conditions with respect to that action (even if he was manipulated into being that way), then if he satisfies the same psychological conditions with respect to his own action, he is responsible for his own action.

It is important to clarify exactly what **PCA1** states: Suppose you steal money from an old man at t_1 and at t_2 you are (still) responsible for stealing his money in virtue of some constellation C of psychological states (the unsheddable stealing values that were in the causal etiology of the action, your memory of having done the action, etc.). Now suppose that for some bizarre and almost certainly

sci-fi reason, at t_3 I come to have the same constellation C of psychological states such that, according to the general psychological connectedness account, I am responsible for the theft. If, as we would expect, I continue acting through those psychological states and steal from a different old man at t_4 through the same constellation of psychological states, **PCA1** says I am also responsible for the new theft. Why? I seem to own my action *at least as much, if not more*, than your action, and thus if I am responsible for yours, I should also be responsible for mine. If I somehow acquire your unsheddable desire to steal and a memory of your theft, and then act through that unsheddable desire in the exact same way that you did, there are *two* reasons for thinking that I am *more* responsible for my action than I am for yours. The first is the fact that *I* am the one who is doing the act, rather than it being one that you did. The second is the temporal fact that I am closer to my action in the sense that my act at t_4 happened more recently than both your act at t_1 and my coming to be responsible for your act at t_3 . But while there might be two reasons for accounting for the possibility that I am responsible for my act but not yours, there is no reason I can think of that would justify my having more ownership over your action than mine.¹⁶⁸

According to the general psychological connectedness account of the diachronic conditions for moral responsibility, manipulated Beth is responsible for Chuck's prior murder. According to **PCA1**, if manipulated Beth is responsible for Chuck's prior murder, then manipulated Beth is responsible for her initial acts of murder. Thus, if we endorse the general psychological connectedness

¹⁶⁸ Indeed, the only other asymmetry is the fact that we are evaluating my responsibility for an act that *was done* compared to evaluating my responsibility for acts that I *do*. But this could not be a relevant factor for two reasons. First, it would be incredible if there were such an asymmetry between the criteria for responsibility for 1) what has been done compared to 2) what is done such that I could be more responsible for an action that was done in the past by someone else compared to an action that I do in the present for the same reasons. Second, more technically, we can imagine that after I rob the old man at t_4 , we assess whether I am responsible, at t_5 , for *what I did* at t_4 . This assessment is even more tightly analogous to the assessment of my responsibility for your action: in both cases I am being judged for an action that *was done*, which I have a memory of, and which were done through the exactly same constellation of psychological states. And so if I am responsible for your action, I must be responsible at t_5 for my (prior) theft because I relate to both my action and your action in the exact same way, but my action happens to be one that I actually did, rather than have a false memory of doing. But if I am responsible at t_5 for an action that I did, then I must be responsible for actually doing the action at t_4 under the simple principle that people can't be non-responsible for an action when they do it, and then spontaneously *become* responsible for having done it.

account, we are committed to upholding Beth's responsibility. This result itself should not be surprising. It would seem implausible that when Beth is manipulated to have Chuck's values, she becomes responsible, on the basis of having those values and some fake memories, for a murder that Chuck did some time in the past, but is not responsible for murders that she herself goes on to do. But, as I have noted, the whole point of historicism is to be able to explain why Beth is absolved of her responsibility for murder when she is manipulated to become like Chuck, and so the general psychological connectedness account does not mesh well with historicism, either positive or negative. That is, the most general version of the psychological connectedness account of responsibility, combined with the principle of comparative autonomy, entails that manipulated agents can be responsible for their actions immediately after manipulation. The following three views form a trilemma:

1. The most general version of the psychological connectedness account
2. The principle of comparative autonomy
3. Historicism about moral responsibility

Given the *prima facie* plausibility of the principle of comparative autonomy, this is at least a *pro tanto* reason against the conjunction of the general psychological connectedness account and historicism.

10.3 The Psychological Account and Instant Agents

In so far as the historicist wishes to absolve manipulated Beth of her responsibility for her initial actions, the general psychological connectedness account appears incompatible with historicism. This incompatibility is particularly damning for positive historicism. Historicists, though they will all agree that instant agents are capable of becoming responsible for some of their future actions, will disagree about whether instant agents are responsible for the first few actions they commit depending on whether they are positive or negative (or operant!) historicists. As a reminder, according to positive historicism, an instant agent is not responsible for his initial actions because he has not had a history

in which he was able to take ownership for his values in the requisite way. Let us integrate this with another case we encountered previously:

Coincidental Doug: At t_1 a man named Doug who satisfies all of the non-historical and historical conditions of moral responsibility freely commits a crime: he assaults another person, but shortly after dies in a car accident. At some point in the future, scientists devise a machine that can create instant agents with any range of possible values, beliefs, desires, etc... They simply need to turn the dials on their machine to the relevant setting, and they can create a person who is non-historically identical to any person in history. The machine also has a 'random' setting that randomly permutes all the possible combinations to create a psychological profile. At t_2 the scientists hit the random button, which creates an instant agent—let us call it iDoug—that coincidentally is a non-historical duplicate of what Doug was like shortly after the assault but before he died. iDoug has a full range of (false) memories of committing the assault.

I have previously argued, in Part II, that iDoug is not responsible at t_2 for the assault that Doug committed at t_2 . This non-responsibility judgment is incompatible with the most general version of a psychological connectedness account of the diachronic conditions of moral responsibility, and so let us suspend that judgment momentarily and concede to the diachronic theorist that iDoug is in fact responsible for an act that he didn't commit, but to which his psychology does bear an intimate relation. We have asked (and conceded the answer to) the question of whether iDoug is responsible for past acts. If we bring the historicist into the picture, we can also ask if iDoug is responsible for further immoral acts that he goes on to commit (with Doug's nefarious values).

Given the structural similarity between this case and the one of manipulated Beth, we could immediately show the complication with positive historicism by appealing to a similar principle:

Principle of Comparative Autonomy 3 (PCA3): If an instant agent is responsible for *someone else's* action in virtue of satisfying certain psychological conditions with respect to that action, then if the instant agent satisfies the same psychological conditions with respect to his own action, then he is responsible for his own action.

This would reveal an incompatibility between positive historicism and the general psychological connectedness account (in addition to the incompatibility just discussed) for the reason that positive historicism denies that instant agents are responsible for their incipient actions, and according to **PCA3** iDoug, an instant agent, is responsible for his.

But perhaps we are less confident about such intuitions about instant agents. We can still reach

the conclusion yielded by **PCA3** by instead considering:

Principle of Comparative Autonomy 2 (PCA2): If a person is manipulated but is still responsible for some action in virtue of satisfying certain psychological conditions with respect to that action, then if an instant agent is created to be a non-historical duplicate of that person and so satisfies the same psychological conditions with respect to that action, then the instant agent is also responsible for that action.

PCA2 states that a manipulated person is as responsible, if not less, for his own action compared to a relevantly similar action done by an instant agent. The justification for **PCA2** is structurally similar to the justification for **PCA1**. If we compare any action done by an instant agent to an act done by a person who is manipulated to be that instant agent's non-historical duplicate and consequently does the exact same action, there are reasons for thinking that the manipulated agent is *less* responsible than the instant agent (because he was robbed of his moral personality and given a new one, whereas the instant agent was just given one). But there is nothing about the instant agent's circumstance that should make us think that he has less autonomy than the manipulated agent; since the manipulated agent is in at least one way less free, and in no ways more free, the instant agent is all-things-considered more free than the manipulated agent. Thus, if the manipulated agent is responsible for the action, then so is the instant agent.

The conjunction of **PCA1** and **PCA2** gives us the same result as **PCA3**. **PCA1** tells us that if manipulated Beth is responsible for Chuck's previous murder, then she is responsible for her subsequent murders. And **PCA2** tells us that if Beth is responsible for her subsequent murders, then an instant agent (like iDoug) is also responsible for his subsequent murders. But this means that iDoug, who is an instant agent, is responsible for the assaults he commits *immediately after creation*. Thus, a general psychological connectedness account, paired with **PCA1** and **PCA2**, commits one to the view that instant agents can be responsible for their first free acts, and is therefore incompatible with any view that denies that instant agents can be responsible for their initial free acts. Since positive historicism is one such view, we have an additional incompatibility between the general psychological

connectedness account and positive historicism.

10.4 Non-historicism to the Rescue?

We ran into several difficulties trying to marry historicism and the general psychological connectedness account. What about when we try non-historicism instead? Certainly, the difficulties we encountered in the last two sections dissolve. According to the psychological account, manipulated Beth and iDoug are both responsible for their relevant predecessor's actions. And given that we hold **PCA1**, **PCA2**, and **PCA3** (whose plausibilities seem independent of which side one falls on the historicism/non-historicism dialectic), they are also responsible for their own actions. But unlike with historicism, this is perfectly kosher under non-historicism, since non-historicism will uphold the responsibility of instant agents and manipulated agents so long as they satisfy the requisite (non-historical) structural criteria for ownership, regardless of how they came to be that way.

Indeed, it is the pairing of general psychological connectedness as the diachronic condition and non-historicism as the synchronic condition that both Matheson and Khoury defend.¹⁶⁹ Both do so by pointing out how the general psychological account helps alleviate some pressures on non-historicism. Both authors note that in cases like manipulated Beth where the manipulation is undone, we seem to have strong intuitions that the “restored” agent is not responsible for the nefarious acts done during the manipulation period. Since we are naturally inclined to think about diachronic conditions for moral responsibility as just a matter of personal identity, the innocence of the restored agent extends to that of the manipulated agent, meaning that manipulated Beth would not be responsible for her crimes. But, and this is the conceptual payload, if we adopt a psychological connectedness view of the diachronic conditions, in lieu of the personal identity view, then our adamancy of restored Beth's innocence need not infect that of manipulated Beth since the two could be psychologically quite dissimilar, particularly with respect to the engineered values that were

¹⁶⁹ Khoury and Matheson (2018). Also see Khoury (2013).

implicated in manipulated Beth's crimes. So the psychological connectedness account dissolves one of the objections to non-historicism.¹⁷⁰ The pairing also seems independently motivated, because if you invoke no causal constraints on responsibility for what was done with the general psychological connectedness account, it seems natural to invoke no causal constraints on responsibility for what is done with non-historicism.

However, while both point out how some problems dissolve when pairing the two views, neither Matheson nor Khoury focus on problems that might be generated for the proponent of the conjunction of the two views. I believe a severe complication for the conjunction of the general psychological connectedness account and non-historicism arises when evaluating the capacity of such a proponent to deal with cases in which agents seem to be responsible for actions when *other people*, but not *they themselves*, do them. It is to this more serious objection that we now turn.

10.5 Moral Scapegoating

Let us suppose that the general psychological connectedness account is true, and roll with the intuition that as a person becomes less and less psychologically connected with the relevant subset of evaluative elements implicated in some prior free action, he loses responsibility for the act. In the case of Leon, after 500 years separate him and his ruthless past, he loses responsibility for his previous malice. Or, to take an actual case, perhaps we find that Klaus Barbie, an elderly man so psychologically distant from the cruel propagator of Nazi ideology in the 40s, is no longer responsible for ordering the deaths of thousands of children. Of course, upholding the general psychological connectedness account does not commit us to defending Barbie's innocence—only that our criterion for determining it hinges on the qualitative similarity of his current psychology with his prior one. It could turn out that when we assess his evaluative elements, we find that he does share enough similarity to his past self to remain

¹⁷⁰ Albeit not the main one (which is the fact that manipulated Beth, even if her innocence is never restored, is responsible for her crimes in the first place).

responsible for the act. This underscores the fact that, even if we know the general psychological connectedness is true, it remains an open question of *what type of psychological connectedness must obtain* for responsibility for some past act to persist or be lost (at a level down from simply the relevant subset).

Rather than try to articulate a substantive position on what sort of psychological connectedness between two agents would constitute the *most plausible version* of the general psychological connectedness account, I wish to instead ask the easier question of what sort of psychological connectedness would constitute an *implausible* version of the account. Obviously, an account that hinged on the psychological connectedness between two agents' desires for chocolate would be an implausible way of cashing out the account, but also one which has no implications for the overall plausibility of meshing non-historicism with the general psychological connectedness account. More interestingly, consider two variants of the Leon case:

Leon 1: Leon at age 30 (Leon-30) – the age he drinks from the well – has the psychological profile of a ruthless conquistador. Imagine that Leon-30 commits a reprehensible act M – namely the massacre of the island's native inhabitants, and satisfies the synchronic conditions of blameworthiness for doing so. Now fast-forward 500 years. Leon is now 530 years old. He has gone through many characters during his life, but with respect to the values permitting the massacre of island natives, the desire to impose his power, and other evaluative elements implicated in act M, Leon-530 remains psychologically connected. Leon-530 shares *almost all* of the distinctive psychological features implicated in act M by Leon-30.

Leon 2: Leon* is in every way identical to Leon in *Leon 1*, with Leon-30* committing act M, but Leon*-530 (temporarily) lacks some executorial capacity that comprises the (synchronic) non-historical conditions for moral responsibility; for instance, he has a brain condition which is such that while the ordinary processes of deliberation and deciding are perfectly intact, when he actually tries to act his motor cortex is non-functional and his ultimate actions are largely random with respect to his all-things-considered judgment and decision.

Before getting to the purpose of this pair of thought experiments, it is useful to flag several things that seem true. First, Leon-530 is responsible for Leon-30's act. The case is designed such that Leon-530 still largely shares the same value system as Leon-30, and so ignoring the intervening life events that were not sufficient to alter his psychology substantially, the only thing that has really changed from a clearly responsible Leon shortly after the murder (say Leon-32) is clock time, which is not morally significant. Second, Leon-530* is responsible for Leon-30*'s act, because it shouldn't matter that he

has lost some capacity that could be relevant to acting now; arguably, when we fall asleep we lose that capacity and yet our putative responsibility for past acts does not change, and Leon-530* is in a similar predicament. Of course, and this is the third observation, if Leon*-530 *now* followed his values and committed a reprehensible act N that is the same as M he would not be responsible because he lacks some capacity that is necessary is satisfying the *non-historical* synchronic conditions for moral responsibility (Leon-530, who has this capacity, would be responsible for N).

But if all these are true, then something interesting emerges; Leon-530* satisfies the non-historical constraints that make him responsible for a *past action* (and since we are dealing with a general account, there are no further historical constraints); but he does not satisfy the non-historical constraints that would need to be satisfied for him to be responsible for an identical *present action*. The underlying principle here is that if I, in virtue of some capacity, am responsible for committing some act in the present, this does not entail that I must retain that capacity to be responsible for having committed that act in the past. This becomes intuitive upon reflection, because while certain executional abilities are clearly needed in order for responsibility to attach to some person's action, it is not plausible that *retaining* such abilities is needed for responsibility to stick with that person through time.¹⁷¹ And since, according to the general psychological connectedness account, responsibility for past actions can be determined *solely* by examining one's psychological states (independent of considerations of personal identity), we know something else is true: *Theon-530*, an instant agent who is created in the world of *Leon 2* to coincide with and be a psychological duplicate of Leon-530* (and so relevantly psychologically connected with Leon-30*), must share all the same responsibility ascriptions we assign to Leon-530*.

¹⁷¹ To take another example, someone could be blameworthy for having intentionally fed an animal bread that he knew was poisonous (satisfying epistemic non-historical conditions that otherwise would excuse him if he acted in ignorance of the poison), but if he subsequently forgets that the bread was and is poisonous (excusing him from doing a relevantly similar act now), this failure to currently satisfy certain non-historical constraints does not seem to excuse blameworthiness for past action.

But this has seemingly implausible implications. Recall:

Principle of Comparative Autonomy 1 (PCA1): If a person is responsible for *someone else's* action in virtue of satisfying certain psychological conditions with respect to that action which he was manipulated to have, then if he satisfies the same psychological conditions with respect to his own action, then he is responsible for his own action.

In this case, Theon-530 is responsible for Leon-30*'s heinous M in virtue of having certain evaluative elements that are psychologically connected to Leon-30* in the same way that Leon-530*'s evaluative elements are; he shares the values, desires, etc., but lacks the actual executional evaluative element needed to bring the act to its free completion were he to do it himself. But it would be strange if Theon-530 becomes responsible for someone else having done an action that he himself would not be responsible for were he to do it, even when his psychological profile in the two cases is the same. Of course, it is clear that this objection, in so far as it is sound, would also apply as an additional problem for the conjunction of the general psychological connectedness view and historicism. However, it demonstrates that the weird implications can still occur without invoking historical manipulations to undermine the autonomy of some agent.

As long as it is true, and I have provided examples suggesting that it is, that the preserved psychological connections needed to render one responsible for a past action *are not coextensive with those needed to render one responsible for performing an action*, we can generate implausible pairs like those comparing Leon-530* and Theon-530. But more important than presenting obscure and seemingly incompatible commitments for the proponent of the two views, the examples seem to highlight how the general psychological connectedness view *accounts for responsibility in the wrong way*. The view basically says that as long as an act is committed freely by someone, anyone can become responsible for it by becoming sufficiently psychologically similar to the person who did the action, *even if they would not be responsible for a similar action that they themselves were to do*. But do I really own an action more if someone else does it freely compared to if I do it unfreely? In the case above, both Theon-530 and Leon-30* have the exact same psychologies performing N and M respectively. The extra property that

undergirds Theon-530's responsibility for Leon-30*'s act M, but not if he were to do a similar act N himself, is that *Leon-30** did it freely.¹⁷²

This seems to suggest that our accounting for Leon-530's responsibility gets the right answer for the wrong reason. This is because Theon-530 by stipulation is not responsible for N, and if he is not responsible for N it is implausible that he is responsible for Leon-30*'s M. But according to the general psychological connectedness account, Theon-530 *is* responsible for Leon-30*'s M, and the same criterion that makes him responsible makes Leon-530* responsible for Leon-30*'s M, *and* Leon-530 responsible for Leon-30's M. So unless we wish to endorse the view that Theon-530 can have this asymmetric responsibility for another's action but not his own, Leon-530's responsibility cannot be explained solely by his general psychological connectedness. Under an alternative account of the diachronic conditions that invokes the relevance of personal identity, Leon-530 has the additional glue of numerical identity that connects him intimately to his past action in a way that Theon-530 does not, and so even though Leon-530's responsibility according to the generally connectedness account happens to align with the common-sense answer (that he is still responsible), the way in which it does so seems wrong.

¹⁷² The only way to avoid this is to revise the account and instead suggest that the general psychological connectedness account is just about how one psychologically relates to *actions in general* (not just free actions). In this case, it is not essential that Leon-30* committed the massacre freely; so long as Theon-530 has basically the same psychology and identifies with the act itself (i.e. complete connectedness save for the executorial capacity that would allow him to actually do a similar massacre freely) it does not matter whether the act was done freely. If this revision is made, then instead of Theon-530 being responsible for Leon-30*'s act but not responsible for what he did (because his own prior action was done unfreely), he is responsible for both. While this might be a more plausible outcome with respect to **PCA1**, it betrays any plausible understanding of how our responsibility for what we do connects to our responsibility for what we've done. This is because it would mean that if Theon-530 commits a similar massacre, he *does not commit the massacre freely*, but yet *moments after* he is responsible for *having committed the massacre*.

CHAPTER XI: The Reasons-Revision and Synchronic Conditions

11.1 Causation to the Rescue?

On top of problems within the view itself, we have found that when we try to conjoin the general psychological connectedness account with both non-historicism or historicism, further problems emerge. Since any account of the diachronic conditions of moral responsibility must be paired with some account of the synchronic conditions, and since any synchronic account will either be a non-historicist or historicist one, we should be especially pessimistic about the viability of any general psychological connectedness account of moral responsibility. Of course, someone sympathetic to psychological connectedness accounts of the diachronic conditions but persuaded by the problems of coincidental people suddenly being responsible would not have been amenable to a general psychological connectedness account in the first place, and so would be comparatively less interested if what already seemed to be an implausible account became even more implausible when paired with another view. Such a person would be more curious about the potential pairings of causal-constraint versions of psychological connectedness with both historicism or non-historicism, and it is the viability of these matches that we turn to next.

Regarding the prospects of pairing a reasons-revision¹⁷³ of the psychological connectedness

¹⁷³ In this chapter, I will just discuss the plausibility of joining the reasons-revision with the alternative synchronic conditions. This is for the following reason. All of the cases used in discussing the general psychological connectedness account's prospective pairing with synchronic conditions can be transformed somewhat straightforwardly to assess either the wide cause or normal cause variation of the psychological connectedness account (for example, we could transform any case into one evaluating the wide account by adding the qualification that the creation of the person/alteration of the person is causally dependent in some fringe way on the original person's psychological states; or for the normal cause account by also grafting and rearranging the brain tissue). I ignore doing these because I believe many of the complications, though admittedly not all, that arise do so independently of adding these additional qualifications (specifically, the argument around Theon-530 is just as strong when we make his creation causally related to Leon-30*, whether through wide or real psychological connections), and thus the incompatibility of the general psychological connectedness account with both historicism and non-historicism would apply equally to a wide cause or normal cause variation of the psychological connectedness account of the diachronic condition of moral responsibility.

Another exclusion in this chapter is worth mentioning: I am not discussing anywhere the plausibility of modifying diachronic accounts with typically historicist qualifications in an attempt to make them more plausible. That is, one might make a sort of "historicist" version of the general psychological connectedness account by requiring there by no causal relationship between the former and latter state that comprise the connection, but requiring that the latter state becomes a part of the person's psychology through a means that occurs with his consent, does not undermine the authenticity of his evaluative scheme, etc. That this can be done to a seemingly "non-causal" account might seem paradoxical at first, but

account with non-historicism, we can start by noting its conceptual advantages compared to its causal-constraintless cousin. Whereas a proponent of the general account was committed to Theon-530 (implausibly) being responsible for a prior act committed by someone else (Leon-30*) but not a relevantly similar prior act that he committed, if we conjoin the reasons-revision with non-historicism we find that Theon-530 is not responsible for Leon-30*'s massacre M in the first place. While he is psychologically connected to Leon-30* in the same way that Leon-530* is, he is an instant agent whose causal etiology has nothing to do with being receptive to the *reasons* that Leon-30* was the way he was, and so he has not come to own those psychological connections in an intimate enough way to warrant becoming responsible for Leon-30*'s act—he does not satisfy that *historical constraints* of being responsible for an act committed by someone else in the past.

But even though the conjunction of the reason-revision with non-historicism is immune from the exact test case faced by the general psychological connectedness account with non-historicism, someone sympathetic to the objection can rightly suggest that this moves too quickly. After all, the objections against the general psychological connectedness account pairing with both historicism and non-historicism boiled down to cases where the same thing that makes an agent responsible for what another person did is not sufficient to make him responsible for doing a comparable action himself, which seems intuitively wrong according to our principles of comparative autonomy. Couldn't the same sort of objection be made here? Someone pursuing this line would just need to sketch out a case

underscores the fact that the general/wide/normal cause variations refer to the causal relationship *between the former and latter states that comprise the connection*, not on potentially other causal constraints on how the latter state came to be the way it is. This certainly would allow us vindicate some historicist intuitions and aid the defense of aspects of some permutations, but I ignore doing this because the force of the arguments remains the same for the following reason. Either this historicist modification to a diachronic condition (i.e. some psychological connectedness account) is done in a permutation with a non-historicist synchronic condition or a historicist synchronic condition. It makes little sense to do it with a non-historicist synchronic condition, because it would be incredible that there are historical constraints for our responsibility for what *was done*, but not analogous historical constraints for what *is done*. And if we do it with a historicist synchronic position, this will not affect the force of the arguments against the permutation because they arise independently of this qualification (i.e. the Theon-530 case applies to historicism as well, and nothing about Theon-530 makes him, *qua* instant agent, not consent to his creation). So while a typical historicist constraint might increase the inherent plausibility of a general psychological connectedness account or a wide one, it fails to mitigate the complications arising from pairing it with historicism or non-historicism.

where someone, Theon-530*, is qualitatively identical to Theon-530, but rather than being created as an instant agent he *comes* to be qualitatively identical by some reasons-bearing causal influence by Leon-30* (or Leon-30*'s ideas) in which he takes himself to have reasons to form similar values, attitudes, etc. Here, the critic would point out that Theon-530* does satisfy the historical constraints according to the reasons-revision to be responsible for Leon-30*'s prior massacre, but since he, like Theon-530, lacks the executorial capacity necessary for freely committing a relevantly similar massacre, should he do so he would not be responsible.

Since, with a little more effort, we can make Theon-530* responsible for Leon-30*'s action according to the reasons-revision and not just the general account, does this demonstrate that the objections we considered against the general account apply equally to wedding historicism and non-historicism with the reasons-revision? Or, instead, does this suggest that the principle of comparative autonomy might seem plausible initially, but it proves too much and we should reconsider the plausibility of someone being responsible for someone else's action but not responsible for committing a relevantly similar action? To start, it is important to stress that **PCA1** is not a blanket principle for any situation in which you are responsible for an act previously done by someone else but aren't responsible for doing a similar one yourself—it is only in cases where you are responsible for the other person's action *in virtue of satisfying some psychological criteria* that you *also satisfy (or would satisfy) with respect to your own action*. And so in the case of the objection against pairing the general psychological account with non-historicism, the idea was that Theon-530 is responsible for Leon-30*'s act solely in virtue of having the same values, desires, and so forth, and thus satisfies the diachronic criterion. But being in these psychological states is not enough for the synchronic criteria, since he lacks the executorial capacity which he apparently doesn't need to satisfy the diachronic condition. Put simply, the synchronic requirements contain all the diachronic ones, plus a few more, but it seems implausible that responsibility for another person's past actions should be more

permissive in this sense than responsibility for your own, and hence the objection.

But in the case of Theon-530*, it is not true that the (non-historicist) synchronic requirements contain all the diachronic ones, as articulated by the reasons-revision, since there is no non-historicist requirement that one comes to have the values he does because of the reason-bearing causal influence of the past responsible real self. Indeed, since the reasons-revision spells out historical requirements, they cannot be a part of a non-historicist account. So the proponent of the reasons-revision paired with non-historicism can at least defend the conjunctive account while still agreeing with his critic about **PCA1**—he merely has to say that there is responsibility-bearing import in the historical criteria of his account of the diachronic condition that, since it is not a component of his synchronic account, can lead to fringe cases where you are responsible for what someone else did but are not for performing a similar act yourself. This at least seems implicit in a Shoemaker comment we have previously considered.¹⁷⁴

It is important to allow for the possibility of my identification with other agents (even strangers), those who do not share my brain but who nevertheless *have* had a significant causal impact on the formation of my real self, through my interaction with them or their ideas. To the extent that their volitional selves have become mine by this causal route, then, so too have their actions.

Conditional on it being true that I can become responsible for someone's past actions by interacting with their ideas, it is plausible that I undergo that sort of interaction with another's ideas without needing the types of executorial capacities that are necessary for satisfying the synchronic conditions of moral responsibility—interacting with ideas is different from acting on them. Thus, the conjunctive account in question allows for fringe cases like Theon-530* where agents can become responsible for another's past actions (by having causally interacted with the past self in the right way) but not be responsible themselves when doing a similar action. It is a further hurdle for the proponent of the conjunctive account in question to argue that this extra autonomy boost of interacting with another's

¹⁷⁴ Shoemaker (2012), 127.

ideas is so strong as to overcome the *prima facie* implausibility of me being responsible for another's past action but not my own. But to be as charitable as possible to the permutation in question, I am willing to make this concession, and argue that even if this were not a problem to conjoining the reasons-revision with non-historicism, serious complications still arise.

To see why, we need only consider what we've just said. The way in which the reasons-revision combined with non-historicism evades the objections levied against the general account and non-historicism is by positing *historical constraints on the diachronic conditions*. This is of course unsurprising, because causation is necessarily historical, and so requiring some sort of causation makes the proponent of the permutation in question a historicist about responsibility for what was done and a non-historicist about responsibility for what is done. While it sounds odd, this is clearly conceptually possible because between any free act and some point in the future when you are presumably still responsible for it, some time has passed; but there is very little conceptual wiggle room in that intervening period for the reasons-revision non-historicist to tell a compelling story. To see why, consider any vanilla case of an instant agent:

Instant Theon: Theon is an instant agent who is created and immediately presented with an opportunity to kill someone. He does so and satisfies the non-historical criteria.

According to non-historicism, Theon is responsible for the murder. According to any plausible conception of moral responsibility, if Theon freely commits the murder then he is responsible for having committed the murder five seconds after doing so. According to the reasons-revision account of the diachronic conditions of moral responsibility, and in contradistinction to some personal identity account of the diachronic conditions, we cannot explain Theon's responsibility moments after the murder by simply saying that he is the same person as the one who committed the murder, but rather by his psychological connectedness with the person who did the murder (himself) and some causal relationship between that past person (himself) and his current self.

Clearly, if a normal/narrow cause psychological connectedness account of the diachronic

conditions were at play, it would be easy to explain the transferred responsibility from Theon's past to his current self. But according to the reasons-revision, as we flagged earlier, it is incredibly difficult to explain how a person's responsibility for various acts sustains over time, because most of the time it does not seem to be the case that the causal flow from our past to our current selves involves reasons at all.¹⁷⁵ While this is a general difficulty for any defender of the reasons-revision, as mentioned earlier, it becomes particularly difficult when trying to pair the view with non-historicism because the non-historicist deems certain actions free even when the evaluative elements in their causal etiology have no history at all. Instant Theon is poofed into existence satisfying the non-historical criteria for the murder he is immediately given the opportunity to commit, but moments after he commits the murder, while there clearly is a causal connection between his past self that committed the murder and his current self, it is *prima facie* implausible that this normal cause continuation of personal identity is in any way reasons-bearing. So when his evaluative elements persist moments later, it is not true that they are authentic according to the reasons-revision account, and the conjunctive account of the reasons-revision and non-historicism leads to cases in which agents act freely but are not responsible for having done such actions only milliseconds in the future.

11.2 The Reasons-Revision and Historicism

And so we come to the final permutation. I have just argued against the viability of the general psychological connectedness account pairing with non-historicism, the general account pairing with historicism, and the reasons-revision pairing with non-historicism. Can we finally make a match between the reasons-revision and historicism? This last permutation seems immune from all the challenges encountered so far in this chapter. Let us refresh ourselves with the main test cases against each of the other permutations:

¹⁷⁵ Nor could we alternatively say that Theon's initial psychological states are caused to be the way they are for reasons, and he simply retains responsibility by those psychological states persisting through time, since as an instant agent he could have been caused to be the way he is for no reasons whatsoever.

When we were assessing the prospects of the general psychological connectedness account and historicism, according to the conjunction, (implausibly) an agent who is manipulated to be an entirely new person is not responsible for an action that stems from that new psychology, but is responsible if an essentially identical act were committed by someone else. This came out in cases where someone like Beth is manipulated to be non-historical duplicate of a criminal like Chuck, in which case according to historicism she is not responsible for subsequent murders she commits with her new evaluative elements, but *is* responsible for prior murders committed by people like Chuck *because* her newly finagled evaluative elements are psychologically connected with those of other criminals in the past. The current permutation (reasons-revision and historicism) encounters no problems here because the proponent of the two views will also deny that Beth is responsible for the past murders of Chuck on the grounds that while she is psychologically connected those psychological connections didn't come about from some reasons-bearing causal influence by Chuck (so she is responsible neither for Chuck's murder nor her own subsequent ones). On the other hand, if she *did* come to be like Chuck in some reasons-bearing way, the causal effects on Beth's change in psychology would not be responsibility undermining since they go through her evaluative scheme, and she is responsible for *both* Chuck's murder and her subsequent actions.

When we were assessing the prospects of the general psychological connectedness account and non-historicism, according to the conjunction, (implausibly) an agent who does not meet some minor non-historical constraint is not responsible for an action that stems from his psychology, but is responsible if an essentially identical act were committed by someone else. This came out in cases where someone like Theon-530 is psychologically connected with some past murderer, but does not satisfy some non-historical constraint like an executional capacity. Again, the current permutation encounters no problems here because the proponent of the two views will also deny that Theon-530 is responsible for the past massacre on the grounds that while he is psychologically connected those

psychological connections aren't the way they are because of the relevant reasons.

Finally, when we were assessing the prospects of the reasons-revision and non-historicism, according to the conjunction, (implausibly) an instant agent who freely commits an action in fact is not responsible, nanoseconds later, for having done the action. This came out in cases like *Instant Theon* who freely commits some action but the evaluative elements in the action's causal etiology, both at the time of the action and moments later, have not become the way they are through any reasons-bearing causation.

Doesn't this objection also apply when pairing the reasons-revision with historicism as well? It seems to for varieties of historicism that allow for the freedom of the incipient acts of instant agents (negative historicism), but I am willing to concede, to be as charitable as possible, that historicisms that specify positive historical constraints do seem to possess the tools to adequately handle these concerns. To see why, we can note that if the historicist is a positive historicist, then *Instant Theon* is underspecified. Since he commits the massacre shortly after coming into existence, it is very likely that under a standard positive historicist account, he has not had sufficient opportunity to undergo rational reflection or engage with his psychology in such a way that would entail his having taken ownership for the evaluative elements undergirding his massacre. But it could be the case that he has had the opportunity for rational reflection (on a McKenna-style positive historicism) such that he is responsible for the massacre.

The upshot of this is that within this ambiguity lies an ability to defuse the objection that if we invoke a reasons-revision version of the diachronic conditions of moral responsibility, we cannot account for Theon's persisting responsibility moments after the murder. We can make a two pronged reply: The first is that *if* Theon in fact freely commits the massacre M, then according to positive historicism it must be the case that he has had the opportunity to engage in the relevant manner with the evaluative elements underlying the massacre (his desire to kill, his lack of concern over others'

lives, etc.)—which is just to say that he meets the *historical constraints* on the synchronic conditions of moral responsibility. And if this is true, then the positive historicist can say that it is true that moments after the massacre, Theon satisfies the historical constraints on the diachronic conditions; his current evaluative elements are the way they are because his past ones were the way they were, and this is because of reasons (Theon’s prior evaluative elements became authentic when he, by stipulation according to positive historicism, had the opportunity to reflect on them, and since his current evaluative elements are the same ones, they also are authentic because he at some point did have the opportunity to reflect on them).¹⁷⁶

On the other hand, *if* Theon did not freely commit the massacre, then since by stipulation he satisfies all the non-historical criteria, it must be because he hasn’t had the opportunity. And if that is true then the mere normal cause persistence of his evaluative elements is not sufficient for them being authentic because such elements need not ever have been reflected on. And so the fact that Theon does not satisfy the historical criteria for the diachronic condition is no longer problematic, because it is only problematic if Theon were actually responsible for the act he had just committed, and on this horn he is not.

To put it another way, the conjunctivist of the reasons-revision and positive historicism has some way of explaining why personal identity seems to matter for past actions, even when it doesn’t. When I freely commit an action, the mere fact that it was freely committed means that I have gone through the right historical process such that the evaluative elements behind the action are authentic. And since those actions are already authentic in virtue of me (at least having the opportunity for)

¹⁷⁶ I myself am not entirely persuaded by this reply, for a similar reason to the one discussed in Part I of this thesis when we looked at the criticism that rational reflection, according to positive historicism, doesn’t actually fashion or shape O-unsheddable values in any meaningful way. If this is true, then it seems like a stretch to say that even if Theon subjects his evaluative elements to rational reflection (and in some sense endorses them *for reasons*) they “are the way they are for reasons”—if the rational reflection couldn’t actually shape the O-unsheddable values, they are the way they are seemingly independent of such reasons. However, as noted, I am willing to grant the soundness of this reply in order to consider the strongest version of the permutation between the reasons-revision and historicism.

rationally engaging with them, and plausibly them being the way they are for reasons, their persistence (through the normal cause way) means that I satisfy the reasons-revision criteria for responsibility for past actions, and we get preservation of responsibility in standard cases of continuity of numerical identity. On the other hand, if someone *else* commits an action, then my mere having psychological connectedness with them is not sufficient for responsibility for their action; I must come to rationally reflect on being that way and so come to endorse or take ownership of those evaluative elements in virtue of some influence they have on me, in which case I can become responsible for another person's actions.

Here it is unsurprising that there are additional constraints on becoming responsible for an act that *someone else* did freely compared to one that I did freely, but importantly we don't need personal identity to explain why in any deep sense. The basic picture is just that you must reflect on your evaluative elements for them to be authentic (or have had the opportunity), and *given that you acted freely*, in so far as in the future your evaluative scheme stays the same, the evaluative elements you now (still) have are guaranteed to be authentic because they are the same ones you previously reflected on. On the other hand, *given that someone else acted freely*, in so far as in the future your evaluative scheme stays the same, the elements you now have are not guaranteed to be authentic because you need not ever have reflected on them.

11.3 The Re-emergence of Re-emergent Autonomy

Unfortunately, the buck stops when we import another objection we previously encountered—though not an objection toward the reasons-revision account, but an objection toward historicism. Consider again, with a slight modification:

One Good Nanosecond: Chuck lives the life of an evil murderer and according to the reasons-revision at some time *t* he is responsible for a murder that was committed in the past. One day manipulators only change Chuck for a nanosecond and then change his modified evaluative scheme back to what it was just prior to the manipulation.

To refresh, in such instances of radical reversals and then radical restorations, it is pretty difficult to

see why the mere happenstances of a manipulation that is immediately undone should in any way prevent a person's responsibility from going on just as if the manipulation had never been done. That is, when Chuck undergoes a radical restoration after a temporary radical reversal, all the things he both was responsible for prior to the manipulation and would be responsible for doing prior to the manipulation should remain that way.

With respect to the actions Chuck goes on to do after his manipulation, the conjunction of the reasons-revision with historicism seems to get on fine. While it's not as simple as non-historicism¹⁷⁷, I previously argued how the historicist—specifically the operant historicist—can note that the manipulations cannot undo what Chuck has done, and in so far as his responsibility is tethered to such practical activity, it remains unchanged between pre-manipulation and post-radical restoration.

However, with respect to the actions Chuck is responsible for prior to his manipulation, this final permutation runs into difficulties. Call Chuck prior to any manipulation pre-Chuck, and Chuck after both manipulations (the reversal and the restoration) post-Chuck. Pre-Chuck and post-Chuck are perfectly psychologically connected with one another, and so the question of whether post-Chuck is responsible for prior actions committed by pre-Chuck depends on the causal connection between the two. But there's the rub: post-Chuck's being the way he is need not depend on pre-Chuck being the way he was, and certainly not for the *reasons* that pre-Chuck was the way he was. Whether or not radical restoration restores what a person was previously responsible for should not depend on the intentions of the manipulators or even their process of achieving that radical restoration, and so in the case where the manipulators randomly manipulated Chuck again (which coincidentally returns him to his pre-manipulation state) there is *no* causal connection between pre-Chuck's being the way he was and post-Chuck's being the way he is. And even if the manipulators deliberately try to return Chuck

¹⁷⁷ Which gets it for free, because as long as the agent was, by stipulation, responsible prior to the very temporary manipulation, undoing that manipulation restores the previous psychological states that already satisfied non-historical criteria.

to his pre-manipulation state, post-Chuck's state certainly hasn't responded to the *reasons* of his taking ownership of his prior state—it is the way it is because of a matching game the manipulators played. So while a general psychological connectedness account of the diachronic condition would allow us to say that post-Chuck (in virtue of being perfectly psychologically connected with pre-Chuck) is responsible for all the things pre-Chuck was, according to the reasons-revision he is not.

We encountered a similar problem to this in the first part of this thesis when evaluating the tools of various camps within the synchronic dialectic to handle the re-emergent autonomy of agents after cases of radical restoration. As a reminder, cases when Chuck (pre-Chuck) is manipulated and then that manipulation is undone (yielding post-Chuck) complicate historicist accounts that require any modifications to evaluative schemes to be authentic for the synchronic conditions of responsibility to be preserved, in so far as one has the intuition that Chuck remains a responsible agent after a large manipulation is undone. The first reply we considered was to divorce the requirement that the historicist conditions be directly connected to his current evaluative scheme; instead of requiring that any relevant values playing a role in an action arose from a history where an agent had the requisite opportunity for rational reflection, the revision requires that any relevant values playing a role in an action are identical to those that arose from a history where the agent (at some point in the past) had the requisite opportunity. We can consider a similar modification to the reasons-revision:

Reasons-Revision: If a person A at t_1 is blameworthy to degree d for act X which occurs at t_1 , then an individual B at t_2 is blameworthy for X to degree d if B is maximally psychologically connected to A through reasons-bearing psychological connections.

Reasons-Revision (modified): If a person A at t_1 is blameworthy to degree d for act X which occurs at t_1 , then an individual B at t_2 is blameworthy for X to degree d if B was at some point maximally psychologically connected to A through reasons-bearing psychological connections in the same way he is now.

In so far as the unmodified reasons-revision requires two things for satisfying the diachronic conditions—sufficient psychological connectivity and the relevant causal history—the modified version allows these two to come apart. I can be causally influenced by another person in such a way

that I become psychologically connected to them and responsible for their past act, but once this reasons-bearing causation has transpired, the psychological connectedness is sufficient for my responsibility; this means that even if I am manipulated and then the manipulation is undone, while my current self (albeit qualitatively identical to myself prior to the manipulation) no longer has the causal connection to the other person, the fact that I did allows me to resume my responsibility when the manipulation is undone. And Chuck is exactly an example of that; pre-Chuck satisfies the reasons-revision for being responsible for some murder that was committed in the past, and according to the modified account so long as Chuck at some point in the past satisfied the reasons-revision account, any point in the future in which he returns to that same psychological connectivity to the past self that committed the murder will restore his responsibility.

When we considered a similar revision toward historicism, the immediate problem was that it was too permissive. Allowing for re-emergent autonomy when manipulations were undone also allowed re-emergent autonomy when *development* was undone, as agents could be reverted to former versions of themselves (say, 10 years prior to any manipulation) and also remain free according to the modified account. The modification for the reasons-revision is equally tone-deaf to any distinction between undoing a manipulation (which intuitively restores autonomy) and undoing natural development (which intuitively does not restore autonomy), because it means as long as I at one point satisfied the reason-revision with respect to responsibility for any past action, I only need general (non-causal) psychological connectedness of the right sort to become responsible again for the past action.¹⁷⁸ Now, the way the historicist got around this issue was by becoming an operant historicist;

¹⁷⁸ While this seems like a compelling objection in the context of historicism (if one is a historicist, rewinding a person's personality to a point in their lives decades ago definitely seems freedom-undermining even if they are similar to who they once were), it is worth pausing for a moment to reflect whether or why this is even a problem for the reasons-revisionist. Consider:

A Former Thief: Tucker is influenced by a thief in a reasons-bearing way such that he develops a desire to steal. He also satisfies the relevant historicist conditions to meet the synchronic requirements for freely stealing, and does so—Tucker is responsible both for the thief's having stolen and for having stolen himself. Fast forward ten

when Chuck is restored to his self just before the manipulation, the evaluative scheme he returns to has elements that are equally operative as the ones in his scheme prior to the manipulation, and so he is immediately eligible for responsibility for subsequent actions. By contrast, if he is reverted to his prior self as it was five years before, the new elements in his evaluative scheme are less operative than the ones they replaced (since he had five years to engage them through practical activity), and so he needs to undergo more practical activity in order to become responsible by acting through them.

Let us consider the viability of a similar move to salvage the reasons-revision. Suppose also that five years prior to his manipulation, Chuck was causally influenced by a thief in a reasons-bearing way such that he developed a value to (in addition to his murderous proclivities) steal, and was sufficiently psychologically connected to the thief that he was, according to the reasons-revision, responsible for the thief's action. In the intervening five years, Chuck becomes psychologically disconnected with the thief and through an authentic process of modifying his evaluative scheme, he abandons the stealing desire. When Chuck is reverted back to some point when he still had the desire, he does not immediately become responsible for it again—it is only after sufficient practical activity that he does. While this gets around the problem, the way it does so paints out a strange picture of

years, and Tucker is a new man who has renounced his old ways and is no longer sufficiently psychologically connected to his past self to be, according to the reasons-revision, responsible for either his or the thief's past acts, and he has an authentic evaluative scheme that lacks the stealing desire. Manipulators come along and revert Tucker's evaluative scheme to when he was a thief, such that he is almost completely psychologically connected with his former self a decade ago. Tucker steals.

According to historicism, this is a canonical example of a freedom-undermining manipulation, and so Tucker is not responsible for his theft, and any historicism that deems him responsible for his subsequent theft fails to vindicate the historicist intuition. But according to the modified reasons-revision, even though Tucker is manipulated out of responsibility for his subsequent thefts, the manipulation restores his responsibility for his past thefts. The reasons-revisionist might point out that this result isn't actually that strange; after all, personal identity is the dominant diachronic condition, in which case once you freely act you're responsible for it forever. Tucker's former thefts, on that view, would travel with him indefinitely, in which case the reasons-revision actually looks more forgiving; compared to the benchmark of Tucker never losing his responsibility for past thefts (if we accept the personal identity account of the diachronic conditions), he does on the reasons-revision and thus the relevant comparison is not that the reasons-revision allows Tucker to be manipulated back into responsibility—on the alternative view, he never gives it up. But this is where the reasons-revision seems to come into tension with the spirit of historicism, for on the face of it, historicism is usually about showing how manipulations strip autonomy, and it is strange that the very same intervention that removes your responsibility for what you do also adds to responsibility for what was done.

moral autonomy. The unmodified reasons-revision already allows the somewhat counterintuitive result of people becoming responsible for others' actions by being influenced by them—but at least there you become responsible for another's action in virtue of some *psychological change in yourself* (that was brought about in the appropriate way). Likewise, in the general psychological connectedness account you become responsible for another's action in virtue of some *psychological change in yourself* (that does not depend on being caused in a particular way).

But on the operativity modification to the reasons-revision under consideration, you can actually become responsible for someone else's action even if there is *no psychological change in yourself*. Here there seems to be a sharp contrast in plausibility compared to how operativity was used in operant historicism. That is because in operant historicism practical activity makes you *eligible* for responsibility for your own actions, even if there is no resulting change in your psychology. This happens all the time, like when we become a member of a club by attending a sufficient number of meetings. On a plausible justification for such requirements, the more we do something, the more we take ownership of it, even if there is no corresponding change to our character.¹⁷⁹ But it is one thing to engage in practical activity to make you responsible for future actions, and quite another to engage in practical activity to make you responsible for past ones. For instance, if a man reads *Mein Kampf* and goes on to discriminate against Jews, it seems extremely bizarre that if he wasn't responsible for some of Hitler's assaults after first reading the book, there is some act of discrimination down the line when his anti-Semitism becomes operative enough such that he is not only responsible for the marginal discrimination but that marginal discrimination *also* makes him responsible for some of Hitler's assaults. *Prima facie*, it is extremely counterintuitive that you could causally influence me and while that wouldn't initially make me responsible for your actions, it could if I engaged with those values enough, and the

¹⁷⁹ I am ignoring the alternative justification for the number of attendances which is that they serve an *epistemic* function, where greater attendances is more reflective of a commitment to join the club, or something along those lines.

dialectical burden is on the reasons-revisionist to explain why this result is not unacceptable.

CHAPTER XII: Personal Identity and the Synchronic Conditions

12.1 Can Identity Fare Better?

In the last chapter, I offered several arguments against all combinations of variations of the psychological connectedness view of the diachronic condition with both historicism and non-historicism. The point was not to proclaim that such pairings are decisively implausible, but to show that the merits of such conceptual marriages are less than the sum of their parts—complications arise in the intersections of positions about the synchronic conditions and positions about the diachronic conditions. In this chapter, I wish to do the same thing toward the other main camp of the diachronic condition—the personal identity view, which says that our responsibility for acts that were done in the past does not depend on our psychological connectedness with past actors, but merely whether we share numerical identity with those prior responsible parties. While the conjunction of the personal identity account of the diachronic conditions and non-historicism runs into just as serious difficulties as discussed with the various psychological connectedness accounts, no obvious ones arise when joining the personal identity account and historicism.

12.2 Personal identity and non-historicism

According to the personal identity account of the diachronic condition, a person is responsible for what someone in the past freely did only if that person is the same person as the free actor in the past. And according to non-historicism, that person is responsible for the actions he commits solely in virtue of satisfying various non-historical requirements. The two views ostensibly are independent, but I hinted earlier at a possible pressure point. Recall, again, Shoemaker's comments on:¹⁸⁰

Gradual Fugue. Suppose that [a fugue] transformation from Johann to Sebastian and back occur gradually (the product of a tumor that grows and then wanes, say) such that psychological continuity is preserved all along. Regardless of such continuity, the Sebastian-actions still do not seem to belong to Johann: while he now remembers them, he nevertheless repudiates them as the actions of someone with a very different psychological make-up and motivational set, actions that seem to Johann as those of a stranger, inexplicable in motivation and justification. Nevertheless, a psychological continuity view of ownership implies that the Sebastian-actions *do* belong to Johann. Psychological continuity is thus insufficient...

¹⁸⁰ Shoemaker (2012), 114-115.

Shoemaker intends to use the example as a counterexample to the view that personal identity is sufficient for moral responsibility. The idea is that the following are true:

1. Johann is not responsible for Sebastian-actions
2. Johann is numerically identical to Sebastian

As I mentioned before, there is an unstated assumption here that is extremely important to flag:

3. Sebastian-actions are done freely

After all, even if personal identity is sufficient for responsibility for past actions, Johann cannot be responsible for Sebastian-actions *if Sebastian-actions are not free*, and the type of fugue state that gradually (and against one's will) replaces his psychological make-up and motivations is *exactly* the sort of intervention that, according to historicism, undermines responsibility. Of course, we are here evaluating the viability of a marriage between the personal identity account and non-historicism. But given that (1), (2), and (3) form a trilemma, paired with the fact that by stipulation Sebastian can satisfy the non-historical criteria for responsibility, there is a problem for the conjunctive account in question. This is because if we stipulate that non-historicism is correct, then (3) is true, and since (2) is true, according to the stipulated truth of the personal identity diachronic criterion, (1) must be false. Thus, in so far as one shares Shoemaker (and my) intuition that (1) is true—that if we are slowly turned into a wildly different person and eventually restored, we are not responsible for those actions that were done when we were so motivationally skewed (even if we freely did such actions)—then there is an incompatibility between the personal identity account and non-historicism.

Nor is Shoemaker's intuition an uncommon one among non-historicists. Khoury, having us imagine a similar situation when an Agent₁ is manipulated to Agent₂, who does something nefarious before autonomy is restored to the reversal agent, Agent₃.¹⁸¹

Responsibility needs to be indexed to time. So when thinking about responsibility in the context of a

¹⁸¹ Khoury (2013), 749.

manipulation case, we need to be clear about the point at which we are assessing responsibility. The question “Is the agent responsible?” is ambiguous. It could mean any of the following:

- (a) Is the pre-manipulation agent (Agent₁) responsible?
- (b) Is the manipulated agent (Agent₂) responsible?
- (c) Is the post-manipulation reversal agent (Agent₃) responsible?

All parties to the debate can agree that (a) and (c) should be answered negatively. It is question (b) that is at issue in the debate between QOW theorists and historicists. The wielders of the manipulation cases often implicitly appeal to the intuitiveness of negative answers to (a) and (c) in order to generate intuitive support for the claim that “the agent is not responsible.” They will often draw attention to the moral innocuousness of the psychology of Agent₁ and Agent₃. Note that in order for it to be plausible both that Agent₂ does something bad and that the pre- and post- manipulation psychology is innocuous the manipulation will have to be severe. That is, if Agent₁ is morally unobjectionable, then since Agent₂ does something seriously wrong (murder, say) then it must be the case that Agent₂ is relevantly psychologically disconnected from Agent₁; similarly for the relation between Agent₂ and Agent₃. In this way, the wielders of the manipulation cases exploit disruptions in psychological connectedness across the manipulations in order to make negative answers to (a) and (c) intuitive. But this is irrelevant since negative answers to (a) and (c) are compatible with an affirmative answer to (b), the question at issue in the debate between QOW theorists and their critics.

Even in Khoury’s framing of the discussion, there is some ambiguity. When we say an agent is responsible, we could mean either that he is a responsible agent in the sense of responsible for the subsequent actions he commits, or responsible in the sense that he is responsible for prior actions. Given that Khoury takes it to be clear that (a) and (c) should be answered negatively, it is clear he is talking about responsibility for Agent₂’s actions.¹⁸² Khoury, like Shoemaker, upholds the moral innocence of the person once his old self has been restored (Agent₃), similarly upholding that the responsibility of Agent₂ does not infect Agent₃’s culpability. And again, Khoury intends to use the thought experiment as a sort of debunking attempt at historicist intuitions (suggesting that historicists only think Agent₂ is non-responsible because Agent₃ is non-responsible and by, mistakenly, believing that the two are symmetrically culpable). But he ignores the implicit result that, since all the agents are the same person (just different time-slices of the person), and given his assertion that Agent₂ is responsible for the malicious act (and responsible for *having done it* for the remainder of time in which

¹⁸² Though this is a strange construction because it is clear that agent₁ is not responsible, since he is around *before* the act in question even occurs!

his psychology is sustained before the manipulation is undone), if Agent₃ is non-responsible for Agent₂'s action then the conjunctive account of personal identity and non-historicism is false.

It is important to clarify the argument I am making. A proponent of the permutation in question might complain in the following way. He would first note that this tension only arises if it is true that, in the case of *Gradual Fugue*, Johann is *not* responsible for Sebastian-actions. While Shoemaker and Khoury, who both have non-historicist leanings, both leave this assertion undefended, they likely do so because they also both explicitly reject personal identity as the right criterion for the diachronic condition; instead they endorse variations of the psychological connectedness view, and so it should be unsurprising that they both have the intuition that in a case like *Gradual Fugue*, where a single person is pushed to become a totally psychologically dissimilar self, a person does not retain responsibility for his acts while in a wildly different state. But someone who endorses the personal identity view will obviously deny that this intuition tracks the truth, because someone who endorses the personal identity view is committed to denying that Johann is not responsible for Sebastian-actions because “Johann” and “Sebastian” both refer to, by stipulation, the same person. So while this is clearly unacceptable to someone with leanings toward the psychological connectedness view, this is a bullet the proponent of the personal identity view and non-historicism must bite. He could then try to explain why this bullet is not impossible to bite, perhaps by noting that a commitment to non-historicism is that even the most seemingly egregious and unwanted manipulations that nevertheless preserve the structural and non-historical requirements of moral responsibility do not undermine one's autonomy. And if I can be radically altered and nevertheless remain free, it seems natural that if I am subsequently radically changed again (and remain responsible for the actions I go on to commit) I retain responsibility for the acts I committed—if radical alterations can sustain responsibility for new actions, it is more than plausible that such radical alternations do not erase responsibility for past ones.

I am not dismissing that such a reply is possible, but pointing out that the fact that such a reply

is needed underscores the relative implausibility of the permutation in question (because other permutations do not need to explain away the intuition). That is, a widely shared intuition is that the following is true:

- (1) Johann is not responsible for Sebastian-actions

I am merely pointing out that if you subscribe to the personal identity view of the diachronic condition, you can deny (1), and if you subscribe to non-historicism you can also deny (1). Remember that there are two ways in which Johann could be non-responsible for Sebastian-actions: either because Sebastian-actions are not done freely, or because, conditional on them being done freely, that responsibility does not transfer to Johann. If you subscribe to the personal identity view, you cannot deny that conditional on Sebastian-actions being free, they transfer to Johann. But you can still deny that Sebastian-actions are done freely; you could do so by asserting historicism. On the other hand, if you subscribe to non-historicism you cannot deny that Sebastian-actions are done freely. But you can deny that conditional on them being free, they do transfer to Johann; you could do so by asserting some psychological connectedness account of the diachronic condition. Thus, in so far as one finds (1) plausible, the personal identity view meshes better with historicism and the psychological connectedness account meshes better with non-historicism. What has additional conceptual costs is the conjunction of the personal identity view with non-historicism.

11.3 Personal Identity and Historicism

And so we arrive at our last permutation. I have argued that neither the general nor reasons-revision of the psychological connectedness criterion of the diachronic condition pairs without complication with either non-historicism or historicism. And I have just argued that the personal identity criterion of the diachronic condition does not pair smoothly with non-historicism. What about with historicism?

The first thing to note is that all the problems that arose from previous permutations of

different synchronic and diachronic criteria can be dodged by the permutation of the personal identity view in conjunction with historicism. The criticism of the general psychological connectedness account with both historicism and non-historicism approximately boiled down to strange implications where people are not responsible for an act that they themselves commit but would be responsible if someone else committed the action. The personal identity view with historicism dodges these merely in virtue of its diachronic commitments; the fact that personal identity is required for responsibility for past actions means that such people who are not responsible for an act that they themselves commit are also not responsible when someone else commits a relevantly similar action. The criticism of the reasons-revision with non-historicism came down to cases where instant agents are responsible for acts they do but not acts they've just done, which the conjunctive account in question obviously avoids because instant agents sustain numerical identity moments after freely acting. So too does the account under consideration avoid the criticism of the reasons-revision with historicism, which had difficulty accounting for the preservation of responsibility when one's psychology changes in a way that seems to preserve responsibility for some acts but such changes are not reason-bearing in nature; so long as it is *one's* psychology that changes, personal identity serves as the glue to conjoin responsibility over time even when psychological changes are not reason-bearing. Lastly, the objection to the pairing of the personal identity account of the diachronic condition with non-historicism led to the difficulty of needing to account for the responsibility of people for the actions they do when in fugue states, which the historicist need not since historicism provides the conceptual resources for saying that such fugue transformations are responsibility-undermining.

In contrast to the other permutations, I cannot conjure up any potential synergistic complications with the permutation of historicism and the personal identity. Obviously, the lack of intersectional difficulties for the conjunction of personal identity and historicism does not on its own provide substantial support for the view—perhaps the inherent complications with each individual

view in the permutation are larger than those of other synchronic and diachronic accounts such that the conjunction is still all-things-considered less plausible than other permutations despite fewer intersectional difficulties. But it does highlight one additional way that the pairing under consideration is more plausible than others, and this is even more so given that *every* other permutation seems to run into difficulties. And the permutation in question does seem independently motivated such that a lack of intersectional problems somewhat vindicates their seeming to be a natural fit; after all, the personal identity account emphasizes the importance of numerical identity for responsibility over time, which historicism nicely echoes by placing constraints on what has happened in *own's own* past. (By contrast, as noted previously, the general psychological connectedness account and historicism pairing seems less motivated because the former gives no historical constraints on one's past for responsibility for what was done, while the latter does give historical constraints on one's past for responsibility for what is done.)

However, there are two larger points that I wish emphasize in closing: the comparative dearth of literature on the diachronic conditions relative to the synchronic conditions of moral responsibility, and the absolute dearth of literature on the intersections between the diachronic and synchronic conditions. Practically speaking, this is problematic because save for the rare instances where we witness atrocities in the making, we only ever blame people for actions that were done in the past. And since blameworthiness for some past action requires both that the action was freely committed and that the person relates to, in the right way, the originator of that action, it is bad practice to put so much conceptual effort into what undergirds responsibility for doing an action, but not responsibility for an act that was freely done. While the long-standing assumption of personal identity comprising the diachronic condition might warrant what would otherwise appear to be philosophical laziness, preliminary work in an underexplored field by writers like Shoemaker, Matheson, and Khoury at the very least suggests that the assumption is less warranted than it is often taken to be.

More so, while I have argued that the alternatives to the personal identity account of the diachronic condition that these authors propose, as well as conceptually possible other alternatives, face complications, even if we end up figuring out that the long-standing assumption is indeed correct, the latter two parts of this thesis hopefully show that further work on the diachronic conditions of responsibility is still a worthwhile endeavor. This is because of the way intersectional conclusions about the diachronic and synchronic conditions provide evidence for the synchronic conditions themselves. For instance, any neutral observer entering the dialectic of the synchronic conditions will agree that either non-historicism or historicism is correct (assuming that moral responsibility *is* compatible with determinism). And since the correct account of moral responsibility comprises of both the synchronic and diachronic conditions—and so there is some correct permutation of the possible synchronic accounts with the possible diachronic ones—if one concluded that the intersectional difficulties between non-historicism and both the personal identity and the psychological connectedness accounts of the diachronic condition were decisive, then if these are indeed the only two plausible accounts of the diachronic condition, then one could conclude that non-historicism is not the right account of the synchronic condition *because* it cannot pair with any plausible accounts of the diachronic condition, and it must to be plausible.¹⁸³ There are a lot of ‘ifs’ here that should make us skeptical of any decisive conclusions we can draw on the basis of the arguments levied so far, but this in no way detracts from the fact that this method provides an additional way of finding evidence against views (like historicism and non-historicism) for which there is a dialectical stalemate in the literature.

This also underscores the attractiveness of such investigations for incompatibilists. While incompatibilists, in believing that moral responsibility is incompatible with determinism, might

¹⁸³ Additionally, if one were persuaded by the first part of this thesis, then the conclusion would be that since either non-historicism or operant historicism is true, given that non-historicism is not, operant historicism is the correct account of moral responsibility.

understandably have little interest in what is the most plausible historicist account of moral responsibility's compatibility with determinism, it is plausible that intersectional difficulties exist and will come to surface for *every* permutation of the synchronic and diachronic conditions, providing evidence that there is no correct account of moral responsibility. Thus, further investigating how the synchronic and diachronic conditions relate might provide new resources for incompatibilist arguments against the possibility of moral responsibility. Perhaps further work, in a similar structure to the arguments I have offered around historicism, will uncover the most plausible diachronic condition of moral responsibility, and the most plausible permutation of the synchronic and diachronic conditions. But perhaps further work will point to the conclusion that there is no correct compatibilist account of moral responsibility, and that in a deterministic world people can never be responsible for their actions. It appears that one way or another, it is in history where we will find the concept of moral responsibility.

References:

- Ayer, A. (1982). "Freedom and Necessity," in A. Ayer, *Philosophical Essays* reprinted in *Free Will*, ed. G. Watson (Oxford: Oxford University Press), 15–23.
- Butler, J. (1975). "Of Personal Identity." In *Personal Identity*, ed. J. Perry (Berkeley: University of California Press), 99-105.
- Campbell, T., & McMahan, J. (2010). "Animalism and the varieties of conjoined twinning." *Theoretical Medicine and Bioethics* 31, 285-301.
- Cappelen, H. (2012). *Philosophy without Intuitions*. Oxford University Press: Oxford.
- Cohen, G. (2006). "Casting the First Stone: Who Can, and Who Can't, Condemn the Terrorists?" *Royal Institute of Philosophy Supplement* 58, 113–36.
- Chalmers, D. (2011). "Verbal disputes." *Philosophical Review*, 120(4): 515-566.
- Chang, R. (2016). "Parity, Imprecise Comparability and the Repugnant Conclusion," *Theoria*, 82, 182-214.
- Christmas, J. (1991). "Autonomy and Personal History," *Canadian Journal of Philosophy* 21(1): 1-24.
- Cuypers, S. (2006). "The Trouble with Externalist Compatibilist Autonomy." *Philosophical Studies*, 129(2), 171-196.
- DeGrazia, D. (2005). *Human Identity and Bioethics*. Cambridge: Cambridge University Press.
- Dworkin, G. (1998). *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press.
- Eshleman, A S. (2001). "Being is not believing: Fischer and Ravizza on taking responsibility." *Australasian Journal of Philosophy* 79, 479–490.
- Fischer, J. (1994). *The Metaphysics of Free Will: A Study of Control*. Blackwell: Oxford University Press.
- Fischer, J. (2014). "Review of *Free Will, Agency, and Meaning in Life*, by Derk Pereboom." *Science, Religion & Culture*, 1(3), 202-208.
- Fischer, J., & Ravizza, M. (1998). *Responsibility and Control: An essay on Moral Responsibility*. Cambridge: 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. (2002). "Reply to John Martin Fischer." In *Contours of Agency: Essays on Themes from Harry Frankfurt*, ed. Buss & Overton (London: MIT Press).
- French, P. (1984). "A principle of responsive adjustment." *Philosophy*, 59(230), 491-503.
- Glannon, W. (1998). "Moral Responsibility and Personal Identity." *American Philosophical Quarterly* 35, 231-249.
- Greene, J., & Cohen, J. (2004). "For The Law, Neuroscience Changes Nothing And Everything." *Philosophical Transactions of the Royal Society B: Biological Sciences*, 1775-1785.

- Haji, I., & Cuypers, S. E. (2004). "Moral responsibility and the problem of manipulation reconsidered." *International Journal of Philosophical Studies*, 12(4), 439-464.
- Haji, I., & Cuypers, S. E. (2007). "Magical agents, global induction, and the internalism/externalism debate." *Australasian Journal of Philosophy*, 85(3), 343-371.
- Haji, I. (1998). *Moral Appraisability: Puzzles, proposals, and perplexities*. Oxford: Oxford University Press.
- Haji, I. (2013). "Historicism, Non-historicism, or a Mix?" *The Journal of Ethics*, 17(3), 185-204.
- Haksar, V. (1980). *Equality, Liberty, and Perfectionism*. Oxford: Oxford University Press.
- Haksar, V. (1991). *Indivisible Selves and Moral Practice*. New York: Oxford University Press.
- Hume, D. (1739)/1978. *A Treatise of Human Nature*, L. A. Selby-Bigge, ed., Oxford: Oxford University Press.
- Khoury, A. (2013). "Synchronic and diachronic responsibility." *Philosophical Studies* 165, 735-752.
- Khoury, A., & Matheson, B. (2018). "Is Blameworthiness Forever?" *Journal of the American Philosophical Association*, 4(2), 204-224.
- King, M. (2013). "The problem with manipulation." *Ethics* 124: 63–83.
- Kolak, D., & Martin, R. (1987) "Personal Identity and Causality: Becoming Unglued." *American Philosophical Quarterly*, 24(4), 339-347.
- Ledoux, S. (2017). "The Barbie Trial of 1987: Narrator of the Holocaust in France." *Revue GenObs* 1(2), 40-71.
- Levy, N. (2011). *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford: Oxford University Press.
- Locke, J. (1975). *Of Identity and Diversity*. In *Personal Identity*, ed. J. Perry (Berkeley, CA: University of California Press), 33-52.
- Madell, G. (1981). *The Identity of the Self*. Edinburgh: Edinburgh University Press.
- Matheson, B. (2014). *Personal Identity and Manipulation Arguments*. Doctoral Thesis: University of Manchester.
- McKenna, M. (2000). "Assessing reasons-responsive compatibilism." *International Journal of Philosophical Studies* 8: 89–114.
- McKenna, M. (2004). "Responsibility and globally manipulated agents." *Philosophical Topics* 32: 169–182.
- McKenna, Michael (2008). "A Hard-Line Reply to Pereboom's Four-Case Argument," *Philosophy and Phenomenological Research* 77/1: 142–59.
- McKenna, M. (2012a). "Defending Nonhistorical Compatibilism: A Reply To Haji And Cuypers." *Philosophical Issues*, 22(1), 264-280.
- McKenna, M. (2012b). "Moral Responsibility, Manipulation Arguments, and History: Assessing the Resilience of Nonhistorical Compatibilism." *The Journal of Ethics*, 16(2), 145-174.

- McKenna, M. (2016). "A Modest Historical Theory of Moral Responsibility." *The Journal of Ethics*, 20(1-3), 83-105.
- McKenna, M., & Coates, D. (2015). "Compatibilism." <https://plato.stanford.edu/entries/compatibilism/>
- McMahan, J. (2002). *The Ethics of Killing: Problems at the Margins of Life*. Oxford: Oxford University Press.
- Mele, Alfred. (1995). *Autonomous agents*. New York: Oxford University Press.
- Mele, Alfred. (2000). "Reactive attitudes, reactivity, and omissions." *Philosophy and Phenomenological Research* 61: 447–452.
- Mele, A. (2007). "Moral responsibility and agents' histories." *Philosophical Studies*, 142(2), 161-181.
- Mele, A. (2008). "Moral Responsibility and History Revisited." *Ethical Theory and Moral Practice*, 12(5), 463-475.
- Mele, A. (2013). "Moral Responsibility, Manipulation, and Minutelings." *Journal of Ethics* 17, 153-166.
- Mellor, D. (1980). *Real Time*. Cambridge: Cambridge University Press.
- Morris, S. (2009). "The Impact of Neuroscience on the Free Will Debate," *Florida Philosophical Review* 9, pp. 56–78.
- Morse, S. (2004). "New Neuroscience, Old Problems." In *Neuroscience and the Law: Brain, Mind, and the Scales of Justice*. New York: Dana Press.
- Parfit, D. (1984). *Reasons and Persons*. Oxford: Oxford University Press.
- Parfit, D. (1986). "Comments." *Ethics* 96, 832-872.
- Parfit, D. (2016) "Can We Avoid the Repugnant Conclusion?" *Theoria* 82: 110–127.
- Pereboom, D. (2001). *Living without free will*. Cambridge: Cambridge University Press.
- Pereboom, D. (2014). *Free will, agency, and meaning in life*. New York: Oxford University Press.
- Pummer, T. (2014). "Does Division Multiply Desert?" *Philosophical Review* 123(1), 43-77.
- Reid, T. (1975). "Of Mr Locke's Account of Our Personal Identity." In *Personal Identity*, ed. J. Perry (Berkeley, CA: University of California Press), 113-118.
- Russell, P. (1988). "Causation, compulsion, and compatibilism." *American Philosophical Quarterly*, 25: 313-321.
- Schechtman, M. (1996). *The Constitution of Selves*. Ithaca: Cornell University Press.
- Shoemaker, D. (2012). "Responsibility Without Identity." *Harvard Review of Philosophy* XXII, 109-132.
- Shoemaker, D. (2014). *Responsibility from the Margins*. Oxford: Oxford University Press.
- Sider, T. (2001). *Four-Dimensionalism*. Oxford: Oxford University Press.

- Strawson, G. (1994). "The Impossibility of Moral Responsibility." *Philosophical Studies* 75(1/2), 5-24.
- Vargas, M. (2005). "The trouble with tracing." *Midwest Studies in Philosophy* 29: 269–291.
- van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Clarendon Press.
- van Inwagen, P. (2000). "Moral Responsibility and Ontology." In *Moral Responsibility and Ontology*, ed. T. van Den Beld (Dordrecht: Kluwer), 11-23.
- Watson, G. (1999). "Soft libertarianism and hard compatibilism." *The Journal of Ethics* 3: 353–368.
- Wiggins, D. (1976). "Locke, Butler and the Stream of Consciousness: And Men as a Natural Kind." *Philosophy* 51(196), 131-158.
- Wilkes, K. (1988). *Real People: Personal Identity Without Thought Experiments*. Oxford University Press: Oxford
- Zimmerman, D. (1999). "Born Yesterday: Personal Autonomy for Agents without a Past," *Midwest Studies in Philosophy* 23: 236–66