8-7-2007

Manipulation and Hard Compatibilism

Daniel Justin Coates

MANIPULATION AND HARD COMPATIBILISM

by

D. Justin Coates

Under the direction of Eddy Nahmias

ABSTRACT

In this paper I consider a recent objection to compatibilism—the manipulation argument. This argument relies on two plausible principles: a manipulation principle that holds that manipulation precludes free will and moral responsibility, and a 'no difference principle' that holds that manipulation is relevantly similar to determinism. To respond to this argument, the compatibilist must reject either the manipulation principle or the 'no difference principle.' I argue that rejecting the manipulation principle offers the compatibilist the most compelling response to the manipulation argument. Incompatibilists claim that this strategy is implausible because it requires that some victims of manipulation are free and responsible. I aim to show that this consequence is not as implausible as it might initially appear.

INDEX WORDS: Free will, Moral responsibility, Compatibilism, Incompatibilism, Manipulation

MANIPULATION AND HARD COMPATIBILISM

by

D. JUSTIN COATES

A Thesis submitted in Partial Fulfillment of the Requirements for the Degree of Master of Arts

In the College of Arts and Sciences

Georgia State University

2007

MANIPULATION AND HARD COMPATIBILISM

by

D. JUSTIN COATES

Major Professor: Eddy Nahmias
Committee: Timothy O'Keefe
Andrew Altman

Electronic Version Approved:

Office of Graduate Studies
College of Arts and Sciences
Georgia State University
August 2007

To my wife,

Stephanie Alice Coates

ACKNOWLEDGEMENTS

This thesis is the result of a number of conversations that I have had with Dr. Eddy Nahmias. In each of these conversations, he carefully listened to my (more often than not) underdeveloped ideas, and in each case, he provided me with helpful feedback and constructive criticisms. I also appreciate my committee members, Dr. Timothy O'Keefe and Dr. Andrew Altman. I am particularly grateful to Dr. John Wingard, who is largely responsible for my pursuit of professional philosophy. I also need to thank Trevor Kvaran, Sean Martin, and Jim Sias for patiently listening to my ideas in their most embryonic form (and for only laughing a little). Finally, I would like to thank my wife Stephanie for her support.

Table of Contents

0. Introduction

Compatibilists believe that (i) the truth of determinism does not preclude free will (FW) and moral responsibility (MR), and (ii) normal adult humans sometimes act freely and responsibly. To argue for their position, compatibilists propose a sufficient condition (or a set of jointly sufficient conditions) for FW and MR that attempts to capture our "inchoate, shared views about [free will] and moral responsibility" (Fischer and Ravizza, 1998; 10). When an agent satisfies this condition (or these conditions), she is free and responsible *even though* such condition(s) can be satisfied in causally deterministic universes. Thus, FW and MR are compatible with determinism. The actual content of these proposed sufficient conditions (PSCs) varies across compatibilist theories.[1] So for example, one PSC might hold that we act freely only if we identify with our desires; while another PSC could hold that we act freely only if our choices accord with what we have best reason to do.[2] Although compatibilists may disagree about the correct content of PSCs, they agree that some PSC (perhaps one that is not currently well-developed) provides a condition on free and responsible agency that (i) satisfies many of our most important intuitions regarding free will and moral responsibility, and (ii) is compatible with determinism.

One important, recent objection to compatibilism is the 'Manipulation Argument.' In this thesis, I will critically evaluate this argument, offering reasons to think that the intuitive plausibility of compatibilism is greater than that of the premises used in the manipulation argument. In §1 I will present the standard version of the manipulation

---

[1] The content varies because as different versions of SC attempt to accommodate our set of shared (Western), inchoate intuitions about FW and MR, each compatibilist is going to find some intuitive aspects of FW and MR to be more important to SC.

Michael McKenna (forthcoming) says that PSCs are a "Compatibilist-friendly agential structure (CAS)," and that they are meant to "exhaust the freedom relevant condition for moral responsibility. Once CAS is satisfied, the agent acts from this structure, allegedly satisfying all that a compatibilist would require for free will."

[2] Other PSCs might deal exclusively with MR (e.g. Fischer and Ravizza's theory of guidance control).

argument. I will also present the content of two types of compatibilist PSCs—structural and historical. I will argue that the standard version of the manipulation argument *apparently* undermines one type of PSC (structural PSCs), but that it is, as it stands, insufficient to topple a second type of PSC (historical PSCs). In §2 I will present two 'cleaned-up' versions of the manipulation argument due to Derk Pereboom and Al Mele respectively. These improved manipulation arguments, I will argue, also offer some *prima facie* reason to reject historical PSCs. In §3 I will argue that these versions of the manipulation argument only offer superficial reasons to reject compatibilism (in either its structural or historical variety). I will argue that in spite of the manipulation argument's rhetorically powerful conclusion, compatibilists have little to fear from the argument itself.

1. Compatibilism and the Problem of Manipulation

1.1 The Standard Argument

The standard version of the manipulation argument relies on the following two principles:

> **Manipulation Principle (MP):** If *S* is manipulated to *A*, then *S* does not freely *A,* and *S* is not morally responsible for *A*.
>
> **No Difference Principle (NDP):** There are not any relevant differences between manipulation and determinism (with respect to FW and MR).

Given these principles, the standard presentation of the manipulation argument (call it the 'standard argument') is best understood as an objection to compatibilism that proceeds as follows: the truth of the conjunction of MP and NDP strongly suggests that if *S* is causally

determined to *A*, then *S* does not freely *A*, and *S* is not morally responsible for *A*. Therefore, compatibilism is false.[3]

But is the standard argument (as it currently stands) successful in showing that compatibilism fails? Compatibilists think not. After all, MP and NDP rely on a problematic characterization of manipulation. Namely, there is no disambiguated notion of 'manipulation' present in MP and NDP (at they are currently articulated), and differences in our understanding of 'manipulation' can affect the plausibility of these principles. To illustrate this problem, let's look at MP.

Incompatibilists typically support MP by appealing to our intuitions in a variety of cases. Consider a paradigmatic case of manipulation, such as a person being hypnotized to cluck like a chicken, and then ask yourself, "is this person freely and responsibly clucking like a chicken?" The quick answer, of course, is "No." So for some paradigmatic cases of manipulation, like hypnosis, MP seems to be plausible. But consider another instance of paradigmatic manipulation, such as a gangster offering a young man $2,500 to murder someone (and we might imagine that the gangster knows that given his desperate financial straights, $2,500 is sufficient to guarantee that the young man will murder the victim). If we consult our intuitions in this case, it is less clear that the young man does not freely murder. In fact, we are probably inclined to think of the young man as free and responsible for his evening of concrete cobbling. Do our intuitions in this case falsify MP? Probably not, but they call for a revision of MP as it currently stands. Thus,

> **MP*:** If *S* is manipulated (in certain ways) to *A*, then *S* does not freely *A,* and *S* is not morally responsible for *A*.

[3] The standard argument represents a family of arguments. For specific examples, see Richard Taylor (1974), Al Mele (1995), Robert Kane (1996), Gary Watson (1999), and Derk Pereboom (2001).

Now, we could spell out the "in certain ways" by providing a list of the particular instances of manipulation that threaten FW and MR. So hypnosis does (threaten FW and MR), and bribery doesn't; guns-to-the-head do, and blackmail doesn't, etc. But this seems like an imprecise measure for whatever explains why manipulation sometimes threatens FW and MR (and why it sometimes does not).[4] So we need an analysis of manipulation that includes a description of when manipulation does, and does not, undermine FW and MR.

Initially, we might be tempted to claim that manipulation threatens FW and MR only if it threatens an agent's ability to satisfy some PSC. But following Michael McKenna, I think that compatibilists should reject this strategy as a "non-starter" (McKenna, forthcoming; 9). What does it mean to reject a strategy as a non-starter? Well, if we were to clarify the content of MP* in terms of inability to satisfy compatibilist PSCs, then we might propose:

> **MP**:** If *S* is manipulated to *A* and because of the manipulation *S* is rendered unable to satisfy some PSC, then *S* does not freely *A,* and *S* is not morally responsible for *A*.

But we should reject this proposal because the sort of manipulation involved in MP** is not relevantly similar to causal determinism. In fact, for compatibilists, they are different in the most important way! That is, causal determinism does not render an agent unable to satisfy compatibilist PSCs, but this sort of manipulation does. And we might charitably assume that no incompatibilist would make *this* mistake. But, according to McKenna, if they do make such a mistake, we should correct it, and suggest an account of manipulation that does not depend on ability (or inability) to satisfy PSCs.

---

[4] To put the point slightly differently, something in the nature of manipulation is such that sometimes we judge that manipulated agents are not free and responsible, and other times we judge that manipulated agents are nevertheless free and responsible, but the concept of manipulation is "wide" enough to accommodate these very different sorts of manipulation.

Does the failure of MP** mean that the standard argument similarly fails? Probably not, but it does indicate that when incompatibilists advance an argument that relies on MP or an MP-like principle, they must clarify what it means to be manipulated, show that causal determinism is relevantly similar to manipulation (in that, agents manipulated in this way are still capable of satisfying some compatibilist PSCs), and offer reasons to think that manipulation, as it appears in their argument, threatens FW and MR—reasons that do not rely exclusively on intuitions about general, paradigmatic cases of manipulation.

But this does not vindicate compatibilism. At best, we should conclude that the standard argument is insufficient to show that compatibilism is false. To succeed, defenders of MP (or some MP-like principle) must provide manipulation cases that do not undermine agents' abilities to satisfy PSCs but that do intuitively undermine FW and MR. Only then can they possibly construct a plausible manipulation principle that, in conjunction with NDP (or some NDP-like principle), suggests the falsity of compatibilism. In what follows, I will consider three manipulation cases. In the first of these cases, the manipulated agent satisfies Harry Frankfurt's PSC, and in the second two cases, we will assume that the manipulated agents satisfy all currently proposed sufficient conditions on FW and MR. But before we consider the merits of these manipulation cases, we must first turn towards an explication of two notable compatibilist PSCs.

1.2 Structural PSCs

Structural theories of free will hold that the PSCs for particular free actions can be met if an agent satisfies such conditions *at a particular time*. So for instance, Harry Frankfurt's account of FW holds that *S* freely wills *A* just in case *S*'s first-order desires (a desire to *x*) mesh with her second-order volitions (a desire to will *x*) at the time of *A* and *S* identifies with *A*

(Frankfurt, 1971). According to Frankfurt, whenever an agent acts, if she satisfies this condition at that time, then she is free. Gary Watson also offers a structural theory of FW (Watson, 1975). Watson suggests that the will is not divided into higher and lower-order desires, but into valuational and motivational systems. Our valuational system tries to answer the question, "what is the best thing for me to do in these circumstances, all things considered?" The motivational system is what moves us to action. According to Watson, when our valuational and motivational systems coincide, we are acting freely in the sense required for moral responsibility. This occurs when what the valuational system determines we should do is also what the motivational system moves us to do. Again, this is a purely structural theory because these conditions—that one's valuational and motivational systems coincide at the time of the action—can be satisfied at a particular instant. Having canvassed two structural PSCs, I will now offer a manipulation case that (i) involves an agent who satisfies these conditions, and (ii) is as "good as it gets" when it comes to eliciting intuitive support for a revised manipulation principle.

St. Patrick's Day Massacre: Suppose that Jones, philosophy's ubiquitous, nefarious neurosurgeon, implants a chip into Smith's brain that causes Smith to violently murder anyone he sees that is wearing a green shirt. This chip makes it such that Smith wants to violently murder everyone he sees wearing a green shirt. It also makes Smith want to will to violently murder all of his town's emerald-clad residents, and identify with that will. On the evening of March 17 at 9:00 P.M., Smith violently kills everyone in his favorite Irish pub. So, at 9:00 P.M. on March 17 Smith's first-order desire matches his second-order volition, and Smith identifies with these desires. According to Frankfurt's structural theory of FW,

Smith is responsible for his St. Patrick's Day massacre.[5] But many will judge that intuitively, Smith *does not* freely massacre dozens of innocents. So satisfying structural conditions at an instant *t* does not seem to be sufficient for FW.

However, we might ask ourselves whether it really is *intuitive* that Smith does not freely massacre dozens. After all, in this case, he *did* want to kill them, want to have the will to kill them, and identify with these desires. The fact that Smith's desires and subsequent identification were manipulated by Smith may just mean that Jones and Smith are *each* blameworthy for the massacre—that each is fully accountable for the tragedy, deserving of our resentment and indignation. In fact, envisioning a scenario such as "St. Patrick's Day Massacre," Frankfurt writes

> There is no paradox in the supposition that [Jones] might create a morally free agent [Smith]. It might be reasonable, to be sure, to hold [Jones] too morally responsible for what [Smith] does, at least insofar as he can fairly be held responsible for anticipating [Smith's] actions. This does not imply, however, that full moral responsibility for those actions may not also be ascribable to the subject (Frankfurt, 1988; 54).

But while Frankfurt seems ready to admit that Smith is free and responsible in "St. Patrick's Day Massacre," many people do not. Frankfurt's response to cases like "St. Patrick's Day Massacre" has been seen as "a hard line indeed" (Kane, 1996; 67). Not surprisingly, upon hearing such a story, many people excuse Smith. That is, they mitigate his responsibility in this case because it seems as if he is just Jones' puppet. This response appears to be natural, and it may seem to many as though the only individuals who do not respond in this natural way are people, like Frankfurt, with an antecedent commitment to the truth of compatibilism.

---

[5] This case can be adjusted slightly to include Watson's PSC. Importantly, the details of such a case would be largely indistinguishable from this case, judgments about Smith's freedom in this case will likely generalize to other structural theories of FW or MR.

1.3 Historical PSCs

In response to such cases, some compatibilists developed *historical* PSCs. Specifically, with compatibilist defenders of historical PSCs on FW and MR, we might think that "an adequate theory of responsibility… must be historical—keeping one eye on the past, to ensure that the actual sequence does not include any responsibility-undermining causes" (Watson, 2004; 211).[6] The "actual sequence" of particular actions, (say, picking a PhD program) goes back, in many cases, past our conscious deliberation (all things considered, should I go to *this* PhD program?) to the mechanisms that produce our actions themselves. For historical compatibilists the genesis or history of these mechanisms matters for FW and MR. According to John Martin Fischer and Mark Ravizza, the "history of (say) an action is important *in part* because it helps to specify what it is for a mechanism to be the *agent's own*" (Fischer and Ravizza, 1998; 170). It is easy to see how understanding FW and MR in historical terms might offer a way for compatibilism to accommodate the purportedly natural intuition that Smith is not free or responsible for the St. Patrick's Day Massacre, while avoiding the unpleasant (at least to compatibilists) conclusion of the manipulation argument. Historical compatibilists can point to manipulation as somehow leading to a failure in the history of a mechanism, as something that prevents the mechanisms that produce our actions from being *our own*. Determinism, it can be argued does not prevent the mechanisms that produce our actions from being our own. This strategy avoids the (to many) distasteful "hard line" response of Harry Frankfurt while preserving our intuitions in cases like "St. Patrick's Day Massacre."

What would a PSC that focused on the historical nature of FW and MR look like? Well, for starters it would include a historical component, as well as something like a

[6] Watson is speaking of John Martin Fischer's PSC on MR—guidance control. See, for example, Fischer 1994 and Fischer and Ravizza 1998.

structural component. For although MR is essentially historical, it is not a *merely* historical phenomenon. That is, although the history of the mechanisms that produce our actions matter (to MR), the immediate source of the action at the time of the action is also important. Fischer and Ravizza offer 'guidance control' as a PSC on MR.[7] Of guidance control Fischer and Ravizza write

> An agent exhibits guidance control of (for example) an action to the extent that the action issues from his own, reasons-responsive mechanism. Thus, there are two important components of this account: the mechanism's being the *agent's own*, and its being appropriately responsive to reasons (1998; 170).

That our actions issue from an appropriately reasons-responsive mechanism fixes the structural component of Fischer and Ravizza's PSC, and that we act from *our own mechanisms* fixes the historical component.

Roughly, to act on an appropriately reasons-responsive mechanism is to act on a mechanism that would do otherwise (i) if there was sufficient reason to do otherwise, (ii) if the agent in question recognized the relevant reason to do otherwise, and (iii) if the mechanism acts according to principled, patterned reasons. So, for instance

> …holding fixed the operation of normal practical reasoning, the pilot [of a commercial airplane] would presumably choose to steer the plane to the east, if told (reliably) that there is a fierce storm to the west (but not to the east). Further, holding fixed the normal, proper functioning of the aircraft (and the lack of a strong wind current), this choice would be translated into action, and the pilot would guide the plane eastward (Fischer, 2006; 18).

But this pilot does not lose this capacity even if causal determinism turns out to be true. She acts on an appropriately reasons-responsive mechanism because if she had had sufficient reasons to do otherwise (there was no storm), she would have likely done otherwise (maintain the same course).

---

[7] For Fischer and Ravizza, guidance control acts as the 'freedom-relevant' condition on moral responsibility. Generally "free will" acts as this requirement, but Fischer and Ravizza seem agnostic about whether "free will" means "the freedom to do otherwise" or something else. If the former, then guidance control takes the place of FW in their account of MR. If the latter, then perhaps FW should be analyzed in terms of guidance control.

Fischer and Ravizza's historical component, an ownership or "taking responsibility" condition, has three components.

1. Individual agents must view themselves as the sources of their actions.[8]
2. Individual agents must recognize themselves to be apt recipients of our reactive attitudes (such as resentment, indignation, gratitude, etc.).
3. "Individual agents' view[s] of [themselves] specified in the first two conditions [must] be based, in an appropriate way, on the evidence" (Fischer and Ravizza, 1998; 213).

Agents satisfy these requirements for 'ownership' by taking responsibility for the mechanisms that produce their actions. This occurs naturally as an agent comes to view herself as a part of the moral community, and as she begins to understand that the mechanisms that produce her actions are causally efficacious (for instance, a young child will discover that she is better able to get what she really wants if she deliberates about what it is that she really wants). When these conditions (both structural and historical) are satisfied, an agent has guidance control over her actions.[9]

Does Fischer and Ravizza's account of MR offer a PSC that can be satisfied by a manipulated agent? To see whether this is possible, we must consider two important versions of the manipulation argument: the Four-case argument and the Zygote argument. These manipulation arguments are improved versions of the standard argument and employ subtly different MP-like and NDP-like principles. If one (or both) of these manipulation arguments can provide a case in which an agent satisfies Fischer and Ravizza's historical

[8] Fischer and Ravizza do not require that agents view themselves as the *ultimate* sources of their own actions—only that they view themselves as the immediate source of their actions, a necessary condition for their own actions to have the sort of causal impact that they do.

[9] This summary of Fischer and Ravizza's account of MR is brief and leaves out many details. However, it is sufficient for our purposes here. For further details see Fischer (1994, 2006) and Fischer and Ravizza (1998).

PSC, yet intuitively, is not responsible for her actions, then compatibilists must (i) reject compatibilism, (ii) abandon Fischer and Ravizza's historical PSC and suggest a PSC that is not subject to this objection, or (iii) attempt to show that the hard line reply is actually attractive (and not counterintuitive). In §2 I will argue that manipulated agents are capable of satisfying historical PSCs (as exemplified by Fischer and Ravizza's account of guidance control). And in §3, instead of jettisoning historical PSCs (or structural for that matter), I will argue that (iii) provides the compatibilist a plausible response to the manipulation argument, as hard compatibilism (of either a structural or historical variety) is superior to its alternatives.

2. Manipulation and History

2.1 Pereboom's Four-case Argument

To begin, Derk Pereboom has developed a four case argument for incompatibilism. These cases feature Professor Plum. Of Plum, Pereboom writes:

> Professor Plum kills Ms. White for the sake of some personal advantage. His…desire to kill White conforms to his second-order desires in the sense that he wills to kill and wants to will to kill, and he wills to kills because he wants to will to kill. In addition, Plum's…process of deliberation is moderately reasons-responsive (Pereboom, 2001; 111).

Notice that Plum satisfies two of the PSCs that we have considered.[10] In the "Four Case" argument, Pereboom provides four different versions of Plum's decision to murder White. These cases lie on a continuum from the covert, local manipulation of each of Plum's actions to natural, causal determinism. Pereboom thinks that our intuitions in the case of covert manipulation should generalize to subsequent cases, and that ultimately, the best

[10] Plum does not currently satisfy Watson's PSC, but it could be easily stipulated that Plum does satisfy Watson's PSC ("Plum judges killing Ms. White the best all things considered, and is thereby motivated to kill Ms. White").

explanation for these intuitions (of Plum's non-responsibility) is that "[Plum's] action results from a deterministic causal process that traces back to factors beyond his control" (Pereboom, 2001; 116). Obviously, if causal determinism is true, then it follows that our actions trace back to factors beyond our control. Thus, the best explanation for our intuitions in cases of manipulation *and* this similarity between manipulation and determinism suggests the falsity of compatibilism. So, let's consider Pereboom's four cases.

> *Case 1*. Professor Plum was created by neuroscientists, who can manipulate him directly through the use of radio-like technology, but he is as much like an ordinary human being as possible, given his history. Suppose these neuroscientists "locally" manipulate him to undertake the process of reasoning by which his desires are brought about and modified—directly producing his every state from moment to moment. The neuroscientists manipulate him by, among other things, pushing a series of buttons just before he begins to reason about his situation, thereby causing his reasoning process to be rationally egoistic. Plum is not constrained to act in the sense that he does not act because of an irresistible desire—the neuroscientists do not provide him with an irresistible desire—and he does not think and act contrary to character since he is often manipulated to be rationally egoistic. His effective first order desire to kill Ms. White conforms to his second-order desires. Plum's reasoning process exemplifies various components of moderate reasons-responsiveness. He is receptive to the relevant pattern of reasons, and his reasoning process would have resulted in different choices in some situations in which the egoistic reasons were otherwise. At the same time, he is not exclusively rationally egoistic since he will typically regulate his behavior by moral reasons when the egoistic reasons are relatively weak—weaker than they are in the current situation (2001; 112-113).

In Case 1, Plum satisfies Fischer and Ravizza's condition of reasons-responsiveness. Nevertheless, we are naturally inclined to think that Plum *is not* morally responsible in this case. Plum is being locally manipulated moment by moment, and intuitively, this seemingly exempts Plum from participating in the moral community. But we might still wonder precisely how covert manipulation has this effect. Pereboom thinks the best explanation for

this exemption stems from all of Plum's decisions being deterministically caused by external forces.[11] To help generate this conclusion, he offers another scenario.

> *Case 2*. Plum is like an ordinary human being, except that he was created by neuroscientists, who, although they cannot control him directly, have programmed him to weigh reasons for action so that he is often but not exclusively rationally egoistic, with the result that in the circumstances in which he now finds himself, he is causally determined to undertake…process…that results in his killing Ms. White. (2001; 113-114).

Again, Plum is covertly manipulated to kill Ms. White, but unlike in Case 1, the manipulation occurs as the result of general programming. Importantly, this general programming doesn't prevent Plum from acting on a reasons-responsive mechanism. Nevertheless, Pereboom claims that intuitively, we should think that Plum is not responsible for killing Ms. White, and again, he thinks that the best explanation for this intuition appeals to deterministic forces that are external to the agent. By comparing Cases 1 and 2, we can see Pereboom's generalization strategy emerge.[12] This strategy takes full shape in Cases 3 and 4.

> *Case 3*. Plum is an ordinary human being, except that he was determined by rigorous training practices of his home and community so that he is often but not exclusively rationally egoistic (exactly as egoistic as in Cases 1 and 2). His training took place at too early an age for him to have had the ability to prevent or alter the practices that determined his character. In his current circumstances, Plum is thereby caused to undertake the…process…that results in his kill White. (2001; 114).

Again, Plum acts on a reasons-responsive mechanism, but unlike in Cases 1 and 2, in response to Case 3, Pereboom does not claim that it is naturally intuitive to hold Plum morally responsible. Instead, he invites the compatibilist to consider her intuitions regarding

---

[11] We might wonder whether the best explanation is not that manipulation *qua* deterministic chain threatens FW and MR, but rather that manipulation *qua* external agent tinkering with my ends (making them his/her ends) is what threatens FW and MR.

[12] A problem for Pereboom emerges here (one that I will not able to fully consider). If Case 2 is considered in isolation (apart from Case 1), it's not clear what our intuitions really our (or should be). Certainly, many compatibilists will think that the fact that Plum was pre-programmed by the neuroscientists to weigh reasons in a particular doesn't preclude his freedom or responsibility. After all, even though the neuroscientists can manipulate Plum to believe that his reasons are good ones for acting, they can't manipulate the justifying relation between good reasons for acting and action, and if Plum acts in accord with those good reasons for acting, then why wouldn't he be free and responsible?

Case 3 (presumably that Plum *is* morally responsible) and offer a principled reason to think that there is a relevant difference between Case 3 and Cases 1 and 2. In the absence of any relevant differences, Pereboom claims that those factors that exempt Plum in Cases 1 and 2 generalize to Case 3. Pereboom concludes this generalization strategy with Case 4.

> *Case 4*. Physicalist determinism is true, and Plum is an ordinary human being, generated and raised under normal circumstances, who is often but not exclusively rationally egoistic (exactly as egoistic as in Cases 1-3). Plum's killing of White comes about as a result of his undertaking the…process[es] in question (2001; 115).

Given Pereboom's claim that Plum is exempt from moral sanction in Cases 1-3, he asks what principled reason compatibilists might have for holding Plum responsible in Case 4. Again, such a reason must offer a relevant difference between Case 4 and the proceeding cases, and in the absence of such a reason, the best explanation for our intuitions is that Plum's non-responsibility in Case 1 generalizes all the way down to Case 4. But if Plum isn't responsible in Case 4, and Case 4 is normal (or possibly normal) given compatibilist assumptions, then no agent is ever responsible in any deterministic scenario. Thus, according to Pereboom, compatibilism is false because no PSCs are actually sufficient for responsibility.

2.2 Mele's Zygote Argument

As if that wasn't enough, the case against historical PSCs can be strengthened. Al Mele invites us to consider the following case of manipulation:

> Diana creates a zygote Z in Mary. She combines Z's atoms as she does because she wants a certain event E to occur thirty years later. From her knowledge of the state of the universe just prior to her creating Z and the laws of nature of her deterministic universe, she deduces that a zygote with precisely Z's constitution located in Mary will develop into an ideally self-controlled agent who, in thirty years, will judge, on the basis of rational deliberation, that it is best to A and will A on the basis of that judgment, thereby bringing about E. If this agent, Ernie, has any unsheddable values at the time, they play no role in motivating his A-ing. Thirty years later, Ernie is a mentally healthy, ideally self-controlled person who regularly

> exercises his powers of self-control and has no relevant compelled or coercively produced attitudes. Furthermore, his beliefs are conducive to informed deliberation about all matters that concern him, and he is a reliable deliberator. So he satisfies a version of my proposed compatibilist sufficient conditions for having freely A-ed (Mele 1995, p. 193).[13]

To help us understand why this case is supposed to be particularly troubling to historical compatibilists, let's unpack Mele's formal presentation of this argument.

(1) Because of the way his zygote was produced in his deterministic universe, Ernie is not a free agent and is not morally responsible for anything.

(2) Concerning FW and MR of the beings into whom the zygotes develop, there is no significant difference between the way Ernie's zygote comes to exist and the way any normal human zygote comes to exist in a deterministic universe.

(3) So determinism precludes FW and MR (Mele 2006, 25).[14]

This case seems so problematic because Ernie lives a life *in the same* way as any other individual. His values and goals are formed through normal processes, and importantly, he can take responsibility for the mechanisms that produce his actions. Whereas Pereboom claims that the falsity of compatibilism is the best explanation for our intuitions in Cases 1 – 4,[15] Mele uses the Zygote case to suggest that compatibilist PSCs can be satisfied by agents who are designed by an external agent. So, manipulation *qua* original design does not preclude an agent's ability to satisfy PSCs. And intuitively, being the product of such a design does threaten FW and MR. Thus, compatibilist PSCs are not actually sufficient for FW and MR.

---

[13] In this presentation of the Zygote case, Ernie does not explicitly satisfy the structural or historical PSCs that we have considered. Ernie does satisfy Mele's own attempt at formulating a compatibilist PSC, but we can stipulate that in addition does satisfying Mele's PSC, Ernie also satisfies the PSCs suggested by Frankfurt, Watson, Fischer, or any other compatibilist for that matter.

[14] This argument is obviously defended by assumptions that *incompatibilists* would draw from the Zygote case. Compatibilists, should (I will argue) draw very different conclusions.

[15] Mele (2006) calls this a 'best explanation' manipulation argument. The premise that the falsity of compatibilism is the best explanation for our intuitions acts as NDP in Pereboom's argument.

2.3 Manipulation and Historical PSCs

Can Fischer and Ravizza's PSC, with its historical component, sidestep the problem as presented by Pereboom and Mele? Initially it may seem so. Remember that according to Fischer and Ravizza, guidance control requires that one takes responsibility or ownership for the mechanisms that produce our actions—it is a historical phenomenon. So in this way, *being morally responsible* is comparable to *being a Goya*. In order to be a Goya, a painting must have the right history. Specifically, it must have been painted by Goya. Similarly, in order for you to be morally responsible for your actions, the mechanisms that produce those actions must be "painted" by you—you must own those mechanisms. So when an agent is manipulated in particular ways (like Plum in Case 1) the mechanism in question is not Plum's own. Rather the mechanism that produces Plum's actions is a conjunction of mechanisms that Plum would take responsibility for as well as the external mechanism (that happens to be inside of Plum's brain) produced and guided by the team of neuroscientists. When Plum, in 'ordinary' circumstances, decides to kill Ms. White, the mechanism that produces this decision and subsequent action is Plum's own (because Plum has taken responsibility for that mechanism). But when Plum is manipulated (as in Case 1), a different mechanism produces the decision to kill Ms. White in Plum. When a skilled forger produces a stroke-by-stroke replica of "The Third of May" the resulting painting is not a Goya because it was not produced by the same mechanism as Goya's "The Third of May." When a skilled manipulator produces an agent with duplicate mental states to a 'free' agent, the action "death of Ms. White" is not free because it was not produced by the same mechanism as that

of the free agent. Furthermore, we might conclude that while manipulation prevents us from acting on mechanisms that we 'own,' determinism does not. [16]

Initially, this seems to be a compelling response. Manipulation, but not determinism, prevents agents from satisfying Fischer and Ravizza's historical requirement because it prevents them from acting on mechanisms that they have not taken responsibility for. But while this response may seem compelling, I believe that it is inadequate. Leaving aside the fact that mechanism individuation raises *generality* problems (McKenna, 2001; Fischer 2004),[17] Fischer's response seems to face another significant problem. To explore this worry, consider what is required for the historical component of guidance control. Agents must view themselves as the (relevant) sources of their actions and as appropriate candidates for reactive attitudes. Moreover, these beliefs must conform (to a large degree) with the evidence an individual has about themselves. Because determinism rules out that we are the ultimate sources of our actions (Kane 1996, Pereboom 2001), Fischer and Ravizza's PSC (which is designed to be compatible with determinism) must employ a notion of 'source' that can be satisfied by a causally determined agent. The agent must see herself as the relevant, proximate source our actions. And, if she consciously controls those mechanisms (to some degree) and they would respond to reasons to do otherwise, then they are 'source-enough.'

But determinism does not rule out this kind of mitigated sourcehood. It does rule out (for humans at least) that an agent can know (or have evidence) for the causal genesis of each of her mechanisms. But knowing this is not required by Fischer and Ravizza's PSC. But this is where the Four-case and Zygote arguments enter into the picture. Envisioning such cases, Gary Watson writes, "why isn't the reasoning that Fischer and Ravizza take to be

---

[16] Fischer (2004) and Fischer and Ravizza (1998) offer this reply (one that turns on our ability to individuate which mechanisms produce an agent's decisions and actions) in response to certain types of manipulation.
[17] These are the 'generality' problems that are similar to those faced by reliablism, rule-consequentialism, Kantianism, etc.

fatal to [structural] theories also fatal to any [historical] compatibilist position, including theirs? Specifically, couldn't the process of taking responsibility be induced by "electronic manipulation of the brain" or some other paradigm responsibility-defeating condition?" (Watson, 2004; 312). Watson then argues that taking responsibility or ownership is compossible with covert manipulation. How? Watson continues

> [Fischer and Ravizza] say that "in taking responsibility for the actions that flow from a kind of mechanism, [one] takes responsibility for acting from the mechanism in all its details" (Fischer and Ravizza, 1998; 216) and emphasize that taking responsibility for a certain mechanism doesn't require knowing "all the details," for example, "the details of the neural states that underlie the mental states that constitute his practical reasoning" (Fischer and Ravizza, 1998; 216). My question is, Why couldn't the details about the exotic origins of the process by among those that one needn't know? According to Fischer and Ravizza, an important feature of the processes that lead to the actions of which me might be ignorant is their "deterministic character." As compatibilists, they think that this ignorance does not rule out our rightly taking responsibility for some of them. Why isn't it just as plausible to think that those meddlesome Martians might have initiated some of the processes for which we rightly take responsibility? (Watson, 2001; 309).

According to Watson, if Fischer and Ravizza's historical component is flexible enough to accommodate an agent genuinely taking responsibility for her causally determined mechanisms, then it is flexible enough to accommodate an agent genuinely taking responsibility for her mechanisms even if she is manipulated by a team of neuroscientists, a deity, or an alien civilization. But this feature of Fischer and Ravizza's PSC generalizes to other historical compatibilist PSCs. So what's a compatibilist to do?

Earlier, I argued that compatibilists have three options. First, they could give up compatibilism. Second, compatibilists could suggest new PSCs that do not seem subject to the worries generated by the manipulation argument. And third, compatibilists could adopt the "hard line" reply of Harry Frankfurt and claim that such a position is not implausible or counterintuitive. In §3 I will argue that compatibilists should opt for the third strategy. I will argue that abandoning compatibilism is not ideal because the alternatives are not without

problems of their own, and further, that the second strategy, called 'soft compatibilism,' ultimately fails. Finally, I will attempt to frame hard line (or hard) compatibilism in novel and attractive ways—ways that will emphasize compatibilism's ability to provide a plausible, intuitive account of many features of human agency without relying on bloated metaphysics.

§3. How to be a Hard Compatibilist

3.1 Why Compatibilism should not be Abandoned

In response to the Four-case and Zygote arguments, compatibilists could reject compatibilism in favor of incompatibilism. But it is not as if incompatibilism is not problematic in various ways. In what follows I will briefly canvass some of the arguments against particular versions of incompatibilism—agent causation, event causation, and hard incompatibilism.[18] The argument that I provide here are not meant to be definitive, but rather suggestive—abandoning compatibilism does not lead us to the promised land. Rather, it leads us to positions that are no less difficult to defend. So, if these arguments are successful, then they provide compelling reasons to not accept incompatibilism. Reasons, as I will argue later in this section, which are stronger than the reasons for rejecting compatibilism on the basis of the manipulation argument.

3.1.2 Problems with Agent Causation

Agent causation, the paradigmatic libertarian conception of human agency, holds that *S* freely wills *A* only if *S* indeterministically causes *A*. Importantly, *S*'s causing of *A* cannot be reducible to physical events occurring in *S*. Rather, it is *S*'s person (self) that causes *A*. So

[18] Obviously, developing serious objections to any of these theories would require a book-length work. I will however, point to problematic aspects of each of these positions, and suggest general strategies that could be used to undermine these positions.

when I decide to type these words, it really is the person Coates that decides to cause them to appear and not the events occurring in my brain which cause me to decide to move my fingers which cause particular movements on the keyboard, etc. Further describing this position, Kane writes

> Libertarian free actions cannot be completely *caused* by prior circumstances, events, or states of affairs; and neither can they be *uncaused* or happen merely by chance…we can say that free actions are indeed caused, but not by prior circumstances, events, or states of affairs. Free actions are caused by the *agent* or *self*, which is not a circumstance, event, or state of affairs at all, but a *thing* or *substance* with a continuing existence…we can say that free actions are *self*-determined or *agent*-caused even though they are undetermined by events. (Kane, 2005; 45).

This picture of human agency seems to capture many of our "inchoate, shared views" about FW and MR. Phenomenologically speaking, it sometimes *feels* as if we are agent causes. So when I make a very important decision, it feels like *I'm* making the very important decision, and that the decision is not reducible to a series of mental or physical events. Moreover, agent causation also seems to make attributions of praise and blame very straightforward. If humans are the unmoved sources of their actions (in the way suggested by agent causation), then clearly, when an individual agent causes some decision or action, there is a clear locus of responsibility—the agent herself. Understanding MR and its accompanying practices this way is quite compelling. But for all of its appeal, agent causation is not without its problems. It seems to conflict with many other domains of human life and enquiry.

Agent causation seems difficult to reconcile to our current scientific picture of the world.[19] As a libertarian account of FW and MR responsibility, agent causation is minimally committed to the following two positions: (i) FW and MR are not compatible with determinism and (ii) normal adult humans sometimes act freely and responsibly. But by committing themselves to (i) and (ii) agent causal theorists (along with libertarians more

[19] For detailed arguments for this conclusion, see Pereboom (2001).

generally) are committed to an a priori denial of causal determinism—an empirical thesis about the underlying physical structure of the universe.

But libertarians do not stop at armchair physics. In order for humans to be agent causes certain facts about human psychology must also obtain. So for instance, the human mind, like George W. Bush, must be the ultimate decider. The mind must have novel downward causal powers (either because mental properties are immaterial or simply properties that emerge from sufficiently complex systems). This understanding of the human mind seems to be metaphysically and scientifically overreaching. It is metaphysically overreaching because it is so demanding—certain supervenience relationships must obtain (or fail to obtain),[20] dualism (in either its substance or property varieties) must obtain,[21] and causation must be transitive.[22] Any one of these highly contentious metaphysical debates could undermine agent causal accounts of FW and MR. But it is also scientifically demanding. As I've noted, it requires that certain physical fact obtain (indeterminism) as well as certain psychological facts (the human mind must be an unmoved source—an admittedly difficult notion to understand). Galen Strawson (1986) has argued against agent causation for philosophical reasons. Being an unmoved mover, or *causa sui*, is nonsense according to Strawson. *Contra* Strawson, Pereboom claims that agent causation is metaphysically possible, but that it is highly unlikely, given our current scientific understanding of the world, that humans have agent causal powers. So we should reject agent causation not for conceptual inconsistencies, but rather, because it is very improbable that we are agent causes. Either way, agent causation seems to have severe difficulties.

---

[20] O'Connor (2000), and O'Connor and Wong (2005)
[21] For a substance dualist agent causal account, see Reid (1788). For a property dualist account, see O'Connor (2000).
[22] See Sartorio (2004) for a sketch of how this objection would go.

In my estimation, these objections should not be used as proofs for the conclusion that agent causation is nonsensical (even though Strawson takes his Basic argument for the impossibility of FW and MR to do so). Rather, they are best thought of as serious problems that agency theorists must be answer. It is important to recognize that no position in the FW and MR debates is uncontested—each position must weigh the costs of its implications. And agent causation, the natural expression of libertarianism, has certain costs, and whether we value a holistic, scientific picture of the world will play a large role in our theory selection.[23]

3.1.2 Problems with Event Causation

Some libertarians are largely in agreement with the critique(s) of agent causation offered in the preceding section. Event causal theorists, like Robert Kane, value a theory of agency that can be reconciled with our current scientific worldview. So, event causal libertarian theories offer an attempt to naturalize libertarianism.[24] According to event causal theorists, if there is the right sort of indeterminism in the causes of my actions (which can be reducible to physical events occurring in my brain), then my actions are free. The sort of indeterminism in question is the same sort posited by current interpretations of quantum mechanics. Thus, when faced with deciding between, for instance, pepperoni pizza and

---

[23] It should be pointed out that many agency theorists are theists, and as such, they don't have the same commitments to a naturalistic worldview. However, I believe, for reasons suggested by Hasker (1989) and Fischer (1994) that libertarian accounts of agency are incompatible with traditional conceptions of God—particularly, God as having perfect foreknowledge of future contingents. These reasons have to do with problems for Ockhamism (see Plantinga 1987), the position that God's knowledge of a future contingent is a soft fact. Some have suggested that Molinism is a way to avoid this problem, but this misunderstands Molinism, which is a position that tries to reconcile libertarian agency with God's providential control. As such Molinism must presuppose Ockhamism (or some other solution to the problem of foreknowledge and freedom). For many theists, revising theologically orthodox positions is not unlike revising scientifically orthodox positions (except in the way such revisions would come about). So if Hasker and Fischer are correct, then even for agency theorists who are not worried about its apparent difficulty to fit with a naturalistic worldview, it does not even fit with the super-naturalistic worldview of many theists.

[24] See Kane (1996), Ekstrom (2000)

supreme pizza, there might be an objective probability of .6 that I will choose the pepperoni and an objective probability of .4 that I will choose the supreme. When I have the right sort of control over my actions (usually the sort of control discussed in compatibilist accounts of agency) and this sort of indeterminism obtains, then I am free and responsible for either of the choices I make.

But if freedom requires this specific type of indeterminism, then event causation is subject to *all* of the criticisms of compatibilism plus it also ties itself to the truth of a particular empirical position (which compatibilism does not). Consider the following manipulation case. Jones implants a chip into Smith's brain that such that the likelihood of Smith killing anyone in a green shirt with his pistol is .6 (Jones didn't like St. Patrick's Day celebrations because he had been pinched as a child) and the likelihood of Smith going quickly home for a quiet evening of Sudoku is .4 (Jones was also the world's largest producer of Sudoku). This chip also ensures that Smith satisfies all the other conditions on freedom and responsibility (be they structural or historical) that compatibilists discuss. But if this is right, then the fact that the quantum events occurred the way they did is outside of Smith's control. And if this is a problem for the compatibilist, then the mere introduction of indeterminism into the system doesn't change that. So the fact that Smith decides to kill all of the pub's green clad patrons is out of his control. By analogy, supposing indeterminism, if you were to flip a coin, there would be an objective probability of .5 that heads would come up and an objective probability of .5 that tails would come up.[25] Even if you really wanted heads to come up, the fact that heads *does* come up, is out side of your control.

[25] It is important not to confuse epistemic probability with metaphysical probability. Even if determinism obtains, a coin flip may still have the epistemic probability of .5 head, .5 tails. But that's not the sort of probability involved in this case.

Similarly, Smith may want to kill all of the pub's patrons, but the fact that the right quantum event occurs in his brain to bring that state of affairs about is beyond his control.

This leaves the event causal theorist in the same position as the compatibilist. Event causation appears to be consistent with certain types of manipulation. Perhaps here the event causal theorist might claim that the above scenario is contrived and might only take up conceptual space (that is, no actual manipulation has ever gone this way). But again, why can't the compatibilist make the same complaint? After all, not many manipulators have ever cared about whether their victims satisfy compatibilist conditions on responsibility. But these sorts of objections seem inappropriate, so I will leave them.

Given that event causation is in the same boat as compatibilism, I think that we should favor compatibilism to event causation because compatibilism accords better with our best science. That is, no matter how our best sciences turn out, event causation of the sort required for freedom and responsibility only obtains if a particular type of indeterminism turns out to be true. So if the physicists were to suddenly say that the world was deterministic (and were right)—or even if the neurobiologists showed that indeterminism does not happen in the right places in the brain—the event causal libertarian would need to revise either her beliefs about the practices and results of science *or* about event causation. On the other hand, compatibilism is compatible with the world having any number of different microphysical theories and neurobiological theories, including the sort of indeterminism (inside or outside the brain) required for event causation. Since we don't know what a completed science will say about whether the laws of nature are deterministic or indeterministic, then we should favor the theory of freedom and responsibility that is consistent with the greatest number of scientific scenarios. And compatibilism fairs better in this way. Again, this is not meant to be a knock-down argument against event causation.

Rather, I am just briefly trying to show that compatibilism is more plausible, all things considered, than event causation. In the next section, I will argue that hard incompatibilism is also less plausible, all things considered, than compatibilism.

3.1.3 Problems with Hard Incompatibilism

Hard incompatibilism is the thesis that incompatibilism is true, but that no one is free or responsible. Traditionally, hard determinism has been the most famous kind of hard incompatibilism, but most contemporary hard incompatibilists eschew hard determinism in favor of a no-free-will-either-way sort of position. Pereboom believes that in order to be free and responsible one must be the ultimate source of one's decisions and actions, but that agent causation is the only way for us to have that power. And since agent causation doesn't seem very likely given our best scientific theories (for reasons I suggested above), Pereboom claims that we should favor hard incompatibilism.

Pereboom offers his Four-case argument to provide an intuitive, compelling case for incompatibilism. He claims that these cases show that compatibilist theories of freedom and responsibility are insufficient because these theories are consistent with covertly manipulated agents being free and responsible. So we should reject compatibilism because it doesn't have the intuitive resources of incompatibilism.

The problem with this approach is that Pereboom prizes our intuitions in a relatively small sample of cases in order to undermine compatibilist PSCs, but then ignores our intuitions a very wide sample of cases once he has established his hard incompatibilism. According to hard incompatibilism, no one is free or responsible. But given cases of ordinary agents, we will certainly judge them to be free and responsible. And certainly this encompasses a much wider array of cases than those cases compatibilists struggle with. In

order to ensure that our intuitions about a few, fanciful manipulation cases are correct, Pereboom has offered a theory that guarantees that our intuitions in almost all normal cases are wrong. But even if the compatibilist must ultimately admit that any agent—even manipulated ones—who satisfies their PSCs is free and responsible, and then attempt to revise our intuitions about such cases (and associated practices), hard incompatibilism certainly leads to more revision. If an "implausible" revision of commonsensical judgments motivates Pereboom's hard incompatibilism, shouldn't the drastic revision of commonsensical judgments implied by his theory similarly motivate compatibilist accounts of agency (assuming that Pereboom and compatibilists are in agreement about the shortcomings of libertarianism)? It seems so.

In this section I have briefly covered some of the major problems with three different types of incompatibilist alternatives to compatibilism. I have not sufficiently shown that any one of these theories is false, but that was never my goal. Rather, I have attempted to show that compatibilists are justified in remaining compatibilists *even if* they do not have a stellar, decisive reply to the manipulation argument. They are justified in remaining compatibilists because to abandon compatibilism would be to adopt a less plausible alternative. Even if there are no good compatibilist replies to the manipulation argument, it does not show that compatibilism is false. Instead, it clarifies the divide between compatibilists and incompatibilists. It, like so many other argument in the FW and MR debates, reduces to a "dialectical stalemate" (Fischer, 1994). But luckily for compatibilism, there are good replies to the manipulation argument. Two types of reply are typically offered—soft and hard. In the remainder of this paper I will argue that we should favor hard replies, and that hard compatibilism is not only plausible, but quite attractive.

3.2 Why Soft Compatibilism doesn't Work

The hard line reply to the manipulation argument (compatibilists who advance such a reply are known as 'hard compatibilists') is characterized by (i) an acceptance of NDP or some other NDP-like principle and (ii) a denial of MP; specifically hard compatibilists will deny that the manipulated agents who figure prominently in incompatibilist stories (Plum, Ernie, etc.) are not free and responsible in those stories.

But MP or some MP-like principle has significant prima facie plausibility. Thus, many compatibilists have opted to (i) accept MP or some suitably disambiguated MP-like principle and (ii) reject NDP. There are, these compatibilists claim, important differences between manipulation and determinism. Because this response seems more palatable to many, it has been called the 'soft line' reply (compatibilists who advance this reply are called 'soft compatibilists').

In many cases, this sort of response seems to work. That is, as I mentioned earlier, many kinds of manipulation are relevantly different from causal determinism. Any instance of manipulation in which the victim is unable to satisfy compatibilist PSCs would be importantly different from determinism. This seems to cover a broad range of the behaviors and activities that are considered manipulation (i.e. hypnosis, Manchurian candidate scenarios, etc.). But although soft compatibilism may seem initially promising, it does not appear to have the resources to deal with cleverer cases of manipulation. Recall that in the Four-case and Zygote arguments the agent in question satisfies many popular compatibilist PSCs (and for any PSCs that are not explicitly mentioned in the scenarios, we can imagine how these sorts of arguments would generalize to those PSCs). In fact, in the Zygote argument, Mele announces that Diana ensures Ernie's future behavior because causal

determinism allows for perfect prediction. So Diana executes her manipulation by using causal determinism. Surely that kind of manipulation is relevantly similar to determinism!

But the soft compatibilist could reply, "perhaps this is correct, and all current compatibilist PSCs can be satisfied by manipulated individuals. That might only show that you are not creative enough in envisioning future compatibilist PSCs that cannot be satisfied by manipulated agents. So, we should work on developing such PSCs, and only take the hard line when such work fails."

In many ways, this response misses the point of my objection. Generals know that at the initial stages of any conflict, you are always fighting the last war. This means that the methods and techniques that served you well in your last war, are the ones that are trotted out in the face of new enemies, and often, these methods no longer work well because of technological advances. The Maginot Line would have served France well in World War I, but it was not particularly effective at stopping the German blitzkrieg of World War II. The same problem arises for the soft compatibilist. Suppose a soft compatibilist develops a PSC that cannot be satisfied by agents who are manipulated in the same ways as Plum or Ernie. This does not end the debate. Likely, an incompatibilist will slightly revise their manipulation case in a few subtle ways that will allow for it to accommodate the new PSC. But then the soft compatibilist will simply add another epicycle to her PSC, and so on. At some point (probably relatively early), we will probably lose sight of whether the PSC in question can accommodate most of our intuitions or whether the sort of manipulation in question really intuitively threatens FW and MR. And there is another worry; if compatibilist PSCs are specifically designed to refute a particular incompatibilist argument, it will be difficult to keep such PSCs from being ad hoc. Similarly, as the cases of manipulation get

more and more watered down to accommodate compatibilist PSCs, it gets less and less clear whether the manipulation in question is even threatening to FW and MR.

So it is not clear that the soft line reply fails in principle, but in terms of dialectical salience, it becomes difficult to see how any soft line reply will provide a satisfying response to a well crafted NDP-like principle. Compatibilists should be hard compatibilists because it is, I will argue, easier to reject MP than NDP. Soft compatibilists, along with incompatibilists, hold that MP is unimpeachable, but this is a mistake. Structural PSCs can be satisfied by manipulated agents. Making compatibilism historical does not help to undermine NDP. So, without some indication of what could threaten NDP, we should favor hard compatibilism. Soft compatibilists and incompatibilists alike think that such a position requires the compatibilist to 'bite the bullet.' But in this case, the bullet is not so hard to swallow.

3.3 Hard Compatibilism

3.3.1 Manipulation and the Meaning of Life

In *The Sirens of Titan* Kurt Vonnegut tells the story of Malachi Constant and Beatrice Rumfoord. Through a series of connected events, Constant and Rumfoord find themselves on Titan, one of the moons of Saturn. Once on Titan, Constant and Rumfoord meet an alien from the planet Tralfamadore named Salo who tells them the true history of Earth. Tralfamadorians, a race of machines, sent Salo on a mission to the other side of the universe. Along the way, some time around 200,000 B.C., an important part of Salo's spaceship broke and he was forced to land on Titan to await the replacement part. Because of the immense distance between Titan and Tralfamadore, it took Salo's request for a new part 150,000 years to reach Tralfamadore. During the last 5,000 years of human history, the Tralfamadorians

used the most powerful substance in the universe to produce in humans the sufficient desires to build certain structures in particular geometric patterns. So

> the meaning of Stonehenge in Tralfamadorian, when viewed from above, is: *"Replacement part being rushed with all possible speed."* …The Great Wall of China means in Tralfamadorian, when viewed from above: *"Be patient. We haven't forgotten about you."* The Golden House of the Roman Emperor Nero meant: *"We are doing the best we can."* The meaning of the Moscow Kremlin when it was first walled was: *"You will be on your way before you know it."* The meaning of the Palace of the League of Nations in Geneva, Switzerland, is: *"Pack up your things and be ready to leave on short notice"* (Vonnegut, 1959; 271-72).

In short, the great achievements of human history, were carefully designed by aliens who were attempting to send messages to a stranded traveler. Civilizations rose and fell simply to ensure that Salo knew when his replacement part would arrive—all under the careful attention of the Tralfamadorians. And the *telos* of human history was a replacement part for a spaceship—a small piece of metal carried by the son of Constant and Rumfoord to Titan!

What if this fanciful story is true of us? Or perhaps, what if instead of an alien race our ends are set by a capricious deity? Or worse (and perhaps more likely), what if instead of being used by aliens or deities, we are just the products of manipulative advertisements that are sufficient to convince thousands to run out an purchase iPhones (even though we have iPods, cell phones, and PDAs already)? Are we not free? Or responsible? How would we respond if we found out that we are subject to such manipulation? Hopefully, we would respond no differently from Malachi Constant who, at the end of his life, in a moment of reflection, said, "it took us a long time to realize that a purpose of human life, no matter who is controlling it, is to love whoever is around to be loved" (Vonnegut, 1959; 313).

This is precisely the hard compatibilist response to the sort of manipulation that is featured in the Four-case and Zygote arguments. There is a real sense in which we set the ends for our lives, and even if an external agent is giving us that end, we can make it our own. We identify with it, we are motivated to seek after it, and we take responsibility for

what drives those ends. In short, we satisfy compatibilist PSCs. But this is precisely where we started; compatibilist PSCs can be satisfied by manipulated agents. But to the hard compatibilist, the appropriate response is not to despair or reject compatibilism, but rather, it is to point out when appropriately disambiguated, such manipulation does not threaten FW and MR. Starting with Malachi Constant and the Tralfamadorians we might ask in what way he was manipulated.

- Did the Tralfamadorians bypass his rational control?
- Did they keep him from identifying with his second-order volitions?
- Did they render him non-reasons-responsive?
- Were their ends irresistible for Constant?

If the answer to any of these questions is 'yes' then the manipulation used by the Tralfamadorians was not relevantly similar to determinism. An argument that relied on this case of manipulation would be a non-starter. If the answer to all of these questions is 'no' then why would such 'manipulation' threaten FW and MR? All the things about manipulation that we fear are listed above. We worry that someone is making us do what we do not want to do. But when the Chinese built the Great Wall, they *wanted* to do it (at least, some of the Chinese wanted to do it). Furthermore, they *wanted to do it for good reasons*—to protect their empire. So they were manipulated to want to do something that would benefit them. What is threatening about that sort of manipulation?

Hard compatibilists say 'nothing' because the sort of manipulation featured in these arguments is not paradigmatic. By this, I mean that paradigmatically, manipulation intuitively threatens FW and MR, and generally it does so in ways that prevent the agent from satisfying compatibilist PSCs. In other words, when manipulation threatens FW and MR it does so in virtue of preventing agents from satisfying compatibilist PSCs. When the

fact that an agent satisfies compatibilist PSCs is made more salient, I suspect that the 'intuitiveness' of Case 1 and the Zygote case will diminish. We could make this fact more salient by reflecting on Plum and Ernie's abilities to control their situation, identify with their ends, and take responsibility for the sources of their actions.

Of course, not all manipulation intuitively threatens FW and MR, and this plays to the hard compatibilist's advantage. Suppose John is manipulated to have the desire to be a good husband and furthered manipulated to want to be the sort of individual that wants to be a good husband (because perhaps, John sometimes finds himself not attending to his marriage). Further suppose that John is manipulated to reflect on these desires, and decides to identify with such desires. Finally, suppose that throughout his adult life, John has been manipulated to take responsibility for the mechanism that produced his identification with those desires. Is John a good husband? Does he deserve praise? It seems to me that he is a good husband and he does deserve praise. Similarly, recall the story of the young man in financially dire straits who is bribed to murder someone for the mob for $2,500. Suppose that given his character, being offered the bribe is sufficient to ensure that he will murder, and also suppose that he identifies with his decision to turn to crime, etc. Is he responsible? I think so. Is the gangster also responsible? Probably so, but that fact alone does not mitigate the young man's responsibility.

There are similar points to be made about Ernie and Plum. Hard compatibilists can rightly challenge the claim that Mele makes in formalizing the Zygote argument: "Because of the way his zygote was produced in his deterministic universe, Ernie is not a free agent and is not morally responsible for anything" (Mele, 2006; 25). Mele argues that intuitively, when we read the Zygote argument, we naturally form this belief. But this is not obvious. Historically, there have always been large populations of individuals whose view of God was

not altogether different from Diana. From Stoics to Calvinists, many individuals would have had the intuition that Ernie *is* free and *is* responsible. Moreover, these traditions traditionally hold that God is perfectly virtuous and *not* responsible for Ernie's behavior. Certainly the intuitions of these individuals are very different from the ones that Mele sought to elicit with the Zygote argument. But how seriously should we take these intuitions?

I do not think that the presence of contrary intuitions shows that Mele's argument does not work, but I do think that it strongly suggests that the support for his MP-like principle is contingent on the cultural and intellectual currents of the day. And recognizing this should lead us to question whether the mechanisms that form our intuitions in any of these fanciful, very contrived cases of manipulation are ultimately reliable. Pereboom's Four-case argument is subject to a similar criticism. An important feature of the Four-case argument is that is proceeds from intuitions in cases of covert, local manipulation to cases of causal determinism. But one might wonder whether it would be more appropriate to proceed from causal determinism to manipulation. After all, if our moral judgment making systems (the systems that regulate ascriptions of praise and blame) 'kick in' in cases that are most similar to how we perceive the world to be, then we should take our intuitions about Cases 3 and 4 to be the more reliable that our intuitions about Cases 1 and 2.

And if this is correct, then the hard compatibilist has already cleared a significant amount of ground. First, she has shown that Pereboom and Mele trade on intuitions about paradigmatic instances of manipulation while advancing cases of non-paradigmatic manipulation. Second, she has suggested that we should be wary about trusting out intuitions in cases of non-paradigmatic manipulation. But there is more to the hard compatibilist project. In the remainder of §3 I will further develop the hard compatibilist position in an effort to make it plausible and attractive.

3.3.2 Manipulation and Universal Exemptions

Peter Strawson claims that there are no universal exemptions on moral responsibility (Strawson, 1962). He distinguished between excusing and exempting conditions. An agent is excused from our practices of praise and blame if she is temporarily unable to satisfy the conditions for freedom and responsibility. So if I have a seizure and that causes me to destroy your brand new plasma television, then you'll be really upset, but you wouldn't blame me (or at least, you shouldn't blame me) because I wasn't in control of my bodily movements. An agent is exempted from our practices of praise and blame if she is not the sort of being that can engage in the moral community. So, dogs, infants, the mentally disabled, etc. are all exempted from our practices of praise and blame because they don't have the general capacities required for engagement in the moral community. Strawson claimed that there could be no universal exemptions because that would undermine the moral community and our practices of praise and blame. Thus, if determinism is true, then it's still appropriate to praise and blame others for their decisions and actions.

But at this point, an incompatibilist could tell the following sort of story: Imagine that just out of telescope range, there is a very powerful and technologically advanced alien civilization monitoring the people on Earth (perhaps Tralfamadore). They notice that people on Earth aren't particularly nice to one another, so they use their very powerful technology to cover Earth with magnetic waves. These waves affect our minds in such a way that we come to want to treat each other nicely, we want to will to treat other nicely, we begin to identify with these desires, these desires and their subsequent actions are produced by a reasons-responsive mechanism, and as we begin to see how much better life is, we take responsibility for these mechanisms. In this way, we have been globally manipulated by the alien civilization to treat each other better, and all the while, we've been satisfying

compatibilist PSCs. In this manipulation case everyone on Earth is affected by the alien waves. We can further stipulate that the aliens have been altering our behavior in this way for hundreds of years, so that the moral community comes to develop around the alien values. So the moral community itself has been manipulated to develop in the way that it does.

What's the moral of this story? Well, I think that it shows that global manipulation can alter the development of the moral community, and so if Strawson is correct about there being no universal exemptions, then it seems that a Strawsonian is committed to the claim that we are morally responsible for the increase in good actions even though we are being subtly manipulated to treat one another better. But this conclusion isn't as implausible as incompatibilists would make it seem. After all, according to Strawson, the moral community shapes the norms of praising and blaming. These practices get their normative depth from the historical features of the moral community and its members (so it's not just the contingent history of *our* moral community that gives our practices their normative depth, but also the capacities and limitations inherent to our humanity). So the Strawsonian can happily admit that global manipulation has occurred and that we are still morally responsible for our decisions and actions. Even this situation should not change our practices of praising and blaming because the practices of the moral community set the norms of praising and blaming. There are no facts about the appropriateness of praising or blaming that independent from the moral community and its inhabitants. And if those facts are also say, Tralfamadorian facts, then so be it. We are still responsible and answerable to one another.

4. Conclusion

Before we get too carried away envisioning ourselves as subjects to an unseen alien overlord, we should remember that there is no good reason to believe that we are globally manipulated. But of course, if we were manipulated in such a way, it would be the hard compatibilist who would have the most satisfying answer. Libertarians and soft compatibilists would lament, "we never had freedom or responsibility." Hard incompatibilists might feel vindicated, "we told you that you never had freedom or responsibility." But hearing that isn't too comforting or satisfying. Only the hard compatibilist can comfort our fears. She can say, "oh, that's an interesting discovery, but I identify with my desires, and I'm motivated by the things I value, and I respond appropriately to reasons (the manipulation never made me do anything compulsive), and I've accepted responsibility for the mechanisms that produce my actions. What else could there be?" And this, I suspect is the right response. Free will and moral responsibility (especially) aren't natural kinds—the facts about these concepts are set not by some outside standard, but rather by our community; they are set by our practice of giving and asking for reasons, holding one another accountable for our actions, praising and rewarding people for doing good, and blaming and punishing people for doing bad. These practices ground our concepts of freedom and responsibility. And insofar as manipulation doesn't affect these practices by preventing us from satisfying some PSC, then we are genuinely free and responsible. And where the manipulation does prevent from satisfying some PSC, then we are not free and responsible. But then again, compatibilists never said we were.

Works Cited

Ekstrom, Laura. 2000. *Free Will: A Philosophical Study*. Boulder: Westview Press.

Fischer, John Martin. 1994. *The Metaphysics of Free Will: an Essay on Control.* Cambridge, MA: Blackwell.

_____. 2004. "Responsibility and Manipulation." *The Journal of Ethics*. 8.2: 145-77. Reprinted in Fischer 2006. Page numbers refer to Fischer 2006.

_____. 2006. *My Way*. New York: Oxford University Press.

_____. and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge University Press

Frankfurt, Harry. 1971. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68: 5-20.

_____. 1988. *The Importance of What We Care About: Philosophical Essays*. New York: Cambridge University Press.

Kane, Robert. 1996. *The Significance of Free Will.* Oxford: Oxford University Press.

_____. 2005. *A Contemporary Introduction to Free Will*. New York: Oxford University Press.

McKenna, Michael. Forthcoming. "A Hard-line Reply to Pereboom's Four-Case Manipulation Argument." *Philosophy and Phenomenological Research*.

_____. 2001. "Review of John Martin Fischer and Mark Ravizza: *Responsibility and Control: A Theory of Moral Responsibility*." *Journal of Philosophy*. 98: 93-100.

Mele, Al. 1995. *Autonomous Agents*. New York: Oxford University Press.

_____. 2006. "Manipulation, Compatibilism, and Moral Responsibility."

O'Connor, Timothy. 2000. *Persons and Causes*. Oxford: Oxford University Press.

_____. and Hong Yu Wong. 2005. "The Metaphysics of Emergence." *Nous.* 39.4: 658-78.

Pereboom, Derk. 2001. *Living Without Free Will*. Cambridge: Cambridge University Press.

_____. 2005. "Defending Hard Incompatibilism." *Midwest Studies in Philosophy: XXIX*: 228-47.

Sartorio, Carolina. 2004. "How to be Responsible for Something without Causing It." *Philosophical Perspectives* 18.

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

Strawson, Peter. 1962. "Freedom and Resentment." *Proceedings of the British Academy* 48: 187- 211.

Taylor, Richard. 1974. *Metaphysics*. Englewood Cliffs: Prentice Hall.

Vonnegut, Kurt. 1959. *The Sirens of Titan*. New York: Delacorte Press.

Watson, Gary. 1975. "Free Agency." *Journal of Philosophy*. 72.8: 205-20.

_____. 1999. "Soft Libertarianism, Hard Compatibilism." *Journal of Ethics*. 3.4: 351-65.

_____. 2001. "Reasons and Responsibility." *Ethics*. 111.1: 374-94.

_____. 2004. *Agency and Answerability*. New York: Oxford University Press.