American Philosophical Quarterly
Volume 55, Number 1, January 2018

THE THREAT FROM MANIPULATION ARGUMENTS

Benjamin Matheson

ABSTRACT

Most seem to presume that what is threatening about manipulation arguments is the "no difference" premise—that is, the claim that there are no responsibility-relevant differences between a manipulated agent and her merely causally determined counterpart. This presumption underlies three recent replies to manipulation arguments. These replies, however, fail to appreciate the true threat from manipulation arguments—namely, the manipulation *cases* that are allegedly counterexamples to the leading compatibilist conditions on moral responsibility. This paper argues that if there is a counterexample to all the leading compatibilist conditions on moral responsibility, then this is sufficient to undermine compatibilism.

1. Introduction

Manipulation arguments are a family of arguments that are typically presented in an effort to undermine compatibilism—the thesis that moral responsibility is compatible with the truth of causal determinism. Each particular manipulation argument is supported by a manipulation case or cases. In most of these cases, we are told that an agent is manipulated by other agents (such as nefarious neuroscientists) to perform some action (normally a morally heinous one). Intuitively, it seems that a manipulated agent is not morally responsible. The catch is that the agent satisfies the leading compatibilist conditions on moral responsibility. Given that the agent seems not morally responsible, it appears that those conditions are, in fact, insufficient for being morally responsible. Incompatibilists then make a further claim that there are no responsibility-relevant differences between such a manipulated agent and a merely causally determined counterpart of hers. Since the manipulated agent seems not morally responsible, it follows that the causally determined agent also seems not morally responsible. Hence, it seems that compatibilism is false.

Most seem to presume that what is threatening about manipulation arguments is the "no difference" premise—that is, the claim that there are no responsibility-relevant differences between a manipulated agent and her merely causally determined counterpart. This presumption underlies three recent replies to manipulation arguments from Kearns (2012), King (2013), and Schlosser (2015). Each argues that manipulation arguments fail because we cannot generalize the nonresponsibility judgment about a manipulated agent to a causally determined agent. Thus, each and every manipulation argument against compatibilism is unsuccessful. But this reply—which I shall call the "no generalization objection"—fails to appreciate the true threat

from manipulation arguments—namely, the manipulation *cases* that are allegedly counterexamples to the leading compatibilist conditions on moral responsibility. I argue that if there is a counterexample to all the leading compatibilist conditions on moral responsibility, then this is sufficient to undermine compatibilism. I show that such a counterexample supports a different argument against compatibilism that I shall call "the control argument."

This paper is structured as follows. In § 2, I outline the general structure of manipulation arguments (what is sometimes called "the" manipulation argument), and I provide an instance of a manipulation argument. In § 3, I outline the no generalization objection. In § 4, I argue that while the no generalization objection undercuts manipulation *arguments*, it does not save compatibilism. This is because, as I show, the real threat from manipulation arguments is not the argument itself, but rather the *manipulation cases* that support them. In § 5, I show that manipulation cases support the control argument.

2. Manipulation Arguments

Manipulation arguments, as noted, are typically deployed in an attempt to show that compatibilism about moral responsibility and causal determinism is false. These arguments share the following structure:

(M1) An agent *S* manipulated in manner *X* to *A* is not morally responsible for *A*-ing.

(M2) There are no responsibility-relevant differences between an agent *S*, manipulated in manner *X* to *A*, and an agent *R* causally determined to *B*.

Therefore,

(M3) *R* is not morally responsible for *B*-ing, even if she satisfies the compatibilist conditions on moral responsibility when she *B*s.

(M4) If compatibilism is true, *R* (if she satisfies the compatibilist conditions on moral responsibility when she *B*s) would be morally responsible for *B*-ing.

Therefore,

(MC) Compatibilism is false.[1]

This is only an argument structure; at this point, it poses no threat to compatibilism. We need another ingredient in order to see the threat for compatibilism—namely, a *manipulation case* to support M1. Such a case describes a scenario in which an agent is manipulated to perform an action (typically a heinous action, such as a murder). Here is one such case:

> [Brainwashed Beth:] Ann is an exceptionally industrious philosopher who works diligently and continuously on being a good teacher, researcher, and colleague. Beth, an equally talented colleague, does not share Ann's devotion to the profession. Beth finds other pursuits more enjoyable and fulfilling, and thus teaches, researches, and does committee work only as much as she must. Their dean wants Beth to be more productive, and so directs a team of psychologists and neuroscientists to figure out what makes Ann tick, and then 'brainwash' Beth so as to make her like Ann. The psychologists determine that it is Ann's 'peculiar hierarchy of values' that makes her so industrious, and the neuroscientists implant the same hierarchy in Beth, while eradicating all competing values. The result is that Beth becomes, in the relevant respects, Ann's psychological twin, now possessing the same industriousness and devotion to her profession. Moreover, the ways in which Ann endorses these values and commitments is now also true of Beth; on critical reflection, they both fully support their ways of life. (King 2013, pp. 68)[2]

Intuitively, Beth seems to lack moral responsibility for her (at least initial) post-manipulation actions. The catch is that Beth satisfies the nonhistorical compatibilist conditions on moral responsibility, such as Frankfurt's (1971) hierarchical conditions when she performs those actions. According to Frankfurt, an agent is morally responsible for an action *A* if her will is properly structured. Roughly, this requires that an agent's

effective first-order desires (the desires that actually move her to action) cohere with her second-order volitions (her desires about which first-order desires she wishes to be effective). Hence, it seems that the above case is a counterexample to Frankfurt's conditions. Now, we might plausibly stipulate that Beth satisfies other leading nonhistorical conditions on moral responsibility. For example, we might stipulate that Beth's action results from a process of deliberation that is reasons-responsive (Fischer and Ravizza 1998), that her reasoning is consistent with her (new) character (Hume 1739/1978, pp. 399–411), that she is sensitive to moral reasons when she acts (Wallace 1994), and that she does not act from an irresistible desire. So this case seems to be (or could become) a counterexample to the leading nonhistorical compatibilist conditions on moral responsibility.[3] So it seems that nonhistorical compatibilism is false.

Incompatibilists aim to take this conclusion further by using manipulation cases like Brainwashed Beth to support a manipulation argument, of the general structure outlined above, to show that compatibilism—and not just nonhistorical compatibilism—is false. Cases like Brainwashed Beth are used to support M1—the claim that a manipulated agent is not morally responsible. Incompatibilists make a further claim when they assert M2: they claim that there are no differences relevant to moral responsibility between Beth and a merely causally determined (i.e., nonmanipulated) counterpart of Beth (call her "Ruth"). If this claim holds, then it follows that compatibilism is false.

3. The No Generalization Objection

McKenna (2008, p. 143) splits responses to manipulation arguments into two kinds: hard-line and soft-line replies. Hard-line replies reject M1. They argue (or sometimes just insist) that, contrary to many people's intuitions, a manipulated agent *is* morally responsible, as long as she satisfies the relevant compatibilist conditions on moral responsibility.[4] Soft-line replies reject M2. They claim that there is a relevant difference between a manipulated and a causally determined agent. They support this claim by identifying conditions on moral responsibility that the manipulated agent has *not* satisfied, but that the causally determined agent allegedly has.[5] McKenna's classification, as we shall now see, is inadequate. There is a response to manipulation arguments that is neither a hard- nor a soft-line reply—namely, the "no generalization objection." This objection is similar to a soft-line reply, as it also, in effect, rejects M2. But, unlike soft-line replies, the no generalization objection does *not* identify further conditions on moral responsibility.

In order to understand how the no generalization objection works, we must first get clear on what is driving our judgments or intuitions about manipulated agents, as this plays a crucial role in the no generalization objection. According to William Lycan (1987/1995, p. 117), "what we object to in these cases is precisely that the victim is the puppet of another person—that his or her *choices* are coerced." So, as Lycan sees it, it is the fact that a manipulated agent is *covertly controlled* (i.e., a puppet of another person) that makes us think she is not morally responsible for her actions. But is he right? It seems that in Brainwashed Beth, for example, there are two senses of "manipulation" at issue:

(1) The neuroscientists covertly manipulate Beth's *brain*.
(2) The neuroscientists covertly manipulate *Beth*.

If (1) is the relevant sense, then Beth is not morally responsible because the neuroscientists covertly manipulate, that is, interfere with, her brain—by changing her brain in some way. If (2) is the relevant sense, then Beth is not morally responsible because the

neuroscientists covertly manipulate *her*—in other words, the neuroscientists covertly control her. Clearly, Lycan holds that (2) is the relevant sense. But we haven't yet ruled out (1) as a contender for what the relevant sense of "manipulation" at issue in Brainwashed Beth is.

It is, in fact, straightforward to rule out (1). If (1) were the relevant sense, then *whatever* sort of brain manipulation an agent undergoes, we ought to find that she is not morally responsible for her subsequent actions. So if neuroscientists were to remove a few of Beth's childhood memories, then we ought to find that she is not morally responsible for her subsequent actions. But this sort of manipulation doesn't seem *necessarily* responsibility-undermining. Suppose Beth, after having a few childhood memories removed, sends a rude e-mail to someone. It's not clear why we ought to think she's not morally responsible for doing so, but that seems to be an implication of (1). Thus, (1) seems like it is not the relevant sense of "manipulation." So the—at least *prima facie* (I'll consider challenges to this claim shortly)—responsibility-undermining factor in a manipulation case is the fact that the agent has been covertly controlled by other agents.

Lycan then claims that we can add a negative condition—a "no covert control" condition—to our analysis of free will/moral responsibility that rules that agents who are the "puppets" of others are not free and so are not morally responsible for their actions. Of course, as Lycan (1987/1995, p. 117) is well aware, such a condition seems "somewhat *ad hoc*." Given that such a condition is ad hoc, it seems that it is no help to compatibilists. It won't help the compatibilist to simply posit a condition that gets them the right result; such a condition must be independently motivated if it is to provide principled grounds for distinguishing between responsible and nonresponsible agents. It seems that such a "no covert control" condition does not provide such principled grounds.

Haji and Cuypers (2001) argue that both compatibilists and libertarians have troubles with manipulation counterexamples, and to avoid such counterexamples, both must endorse what they call a "no bypassing" condition. The "no bypassing" condition is, I think, a form of a "no covert control" condition. Haji and Cuypers implicitly attempt to sidestep the worry that such a condition is ad hoc by claiming that *all* the positions in the debate require such a condition. This move, however, ignores (at least) one position: the impossibilist. The impossibilist holds that no one is morally responsible because the conditions on moral responsibility are impossible to satisfy; for example, they claim a morally responsible agent must be self-creating. Once we include the impossibilist, the "no bypassing" condition—and *a fortiori* the "no covert control" condition—is clearly ad hoc since its only motivation is to get the moral responsibility possibilist (i.e., compatibilists and libertarians) the result they want.

Barnes (2015) and Waller (2014) have recently defended more sophisticated "no covert control"-style conditions (though they don't call them that) that potentially avoid this worry. However, as I argue elsewhere (Matheson, unpublished manuscript; see also Matheson 2016, pp. 1968–1969), there are possible manipulation cases *without* agent-manipulators—that is, cases where an agent is "controlled" by an intentionless force, and yet the agent still seems to lack moral responsibility. Such cases act as counterexamples to the proposed "no covert control" conditions. To avoid such counterexamples, compatibilists would then have to posit a condition that also ruled out "control" by an intentionless force. But such a condition seems patently ad hoc.[6]

While a "no covert control" condition is unpromising, the no generalization objection

is pressed *without* claiming that there is such a condition on free will/moral responsibility. Instead of taking Lycan as proposing such a condition, we can take him as diagnosing the *source* (or primary cause) of the nonresponsibility judgment about manipulated agents—namely, the fact they were covertly controlled. If it is the case that this fact is the source of the nonresponsibility judgment, we have a *prima facie* reason *not* to generalize the nonresponsibility judgment from (say) Beth to Ruth. This effectively stops a manipulation argument against compatibilism supported by Brainwashed Beth—and indeed any manipulation argument against compatibilism—in its tracks.

The no generalization objection has been defended, in various forms, by Kearns (2012), King (2013), and Schlosser (2015).[7] While each of their replies differs in the details, they all make the same core point: namely, there are responsibility-relevant differences between a manipulation case and a mere determination case such that we cannot reliably generalize the nonresponsibility judgment about a manipulated agent to a merely determined agent. The reason is that the nonresponsibility judgment apparently stems from a feature *unique* to a manipulation case—namely, the fact that the manipulated agent has been covertly controlled by another agent. Because this feature is *not* (by hypothesis) present in a mere determination case, there is no ground for generalizing the nonresponsibility judgment from a manipulation case to the mere determination case. Hence, M2 of *any* manipulation argument is false. Therefore, manipulation arguments fail.

Some might worry that commitment to the no generalization objection might amount to or entail commitment to a "no covert control" condition on moral responsibility. I think this is incorrect. The defender of the no generalization objection need not be committed to there being particular conditions on moral responsibility, because this objection only aims to show that manipulation arguments do not undermine compatibilism. The incompatibilist claims that causal determination is the responsibility-undermining factor in the relevant manipulation cases. To defeat the incompatibilist's argument, compatibilists could identify extra conditions on moral responsibility that haven't been satisfied by the manipulated agent in question, but they need not. Instead, they might simply show that causal determination is not the responsibility-undermining factor in these cases. This is what the no generalization objection does. In effect, this shows that manipulation arguments do not foreclose the *possibility* of conditions beyond those currently defended by compatibilists. But there is no need for compatibilists to take on the burden of supplying these extra conditions. Of course, a full defense of compatibilism might require positing such conditions. However, since the no generalization objection's aim is only to undercut an argument against compatibilism, it is not forced to take on the burden of providing further conditions to undercut this argument. The argument is undercut if it can be shown that causal determination is not the responsibility-undermining factor in the relevant manipulation cases.

Patrick Todd (2013) argues against the no generalization objection (though he also doesn't call it that).[8] He argues that manipulation (and presumably, covert control) is irrelevant to what actually *makes* a manipulated agent nonresponsible, even though it might be what produces our intuitive judgment that a manipulated agent is not morally responsible. While the no generalization objection claims that we cannot generalize our *judgment* about a manipulated agent to a nonmanipulated and merely causally determined agent, Todd claims that we actually generalize the *fact* of what makes a manipulated agent nonresponsible—and this, according to Todd, comes

apart from what produces our judgment that a manipulated agent is not free or morally responsible. Todd writes:

> The proponent of [a manipulation] argument should admit that *the manipulation [or covert control] does no work in making the agent unfree [or nonresponsible]*. Rather, the proponent of the argument contends—and clearly must contend—that the *manipulation [or covert control] is irrelevant as concerns what makes the agent unfree*. She instead says that the manipulation can *help us see* that something does make the agent unfree. In other words, she first presents the scenario (say) to an agnostic, and asks whether the agnostic thinks that the agent is free (or responsible) in that scenario. And suppose the agnostic says 'no.' She then points out that whatever would make the agent unfree in that scenario would also make the agent unfree in a qualitatively identical scenario, except in which blind natural causes have taken the place of an intentional agent. (Todd 2013, p. 202; emphasis added)[9]

The idea is that after an agnostic (i.e., someone without a commitment to either compatibilism or incompatibilism) has been presented with a manipulation case, they judge that the featured individual is not morally responsible.[10] So suppose an incompatibilist presents Brainwashed Beth to an agnostic, and the agnostic then judges that Beth is not morally responsible. Todd claims that the incompatibilist can then *tell* the agnostic that the covert control was not, in fact, relevant to why Beth is not morally responsible. The incompatibilist is then able to provide an alternative explanation that is conducive to incompatibilism—for example, Beth is not morally responsible because she has been causally determined by events beyond her control. In effect, it seems that Todd is trying to push through the generalization from a manipulation case to a determination case (such as from Brainwashed Beth to Determined Ruth).

However, by claiming that manipulation (or covert control) is *irrelevant* to what makes a manipulated agent nonresponsible, Todd seems to have undercut our only reason for holding that Beth (or any other manipulated agent) is not morally responsible. We form the judgment that Beth is not morally responsible on the basis that she has been covertly controlled, but then Todd tells us that covert control is irrelevant to what makes her nonresponsible. But once he does that, it's not clear that we have a reason to continue holding that Beth is not morally responsible. If we form a judgment *J* for reason *R*, but then we find out *R* is irrelevant to *J*, then we should reject *J*. And we've been told that manipulation is irrelevant as to what makes a manipulated agent nonresponsible.

Of course, while manipulation might not *make it the case* that Beth is not morally responsible, it might still act as *evidence* that she is not morally responsible.[11] Note, though, this move accepts that manipulation *is* relevant (i.e., epistemically) to what makes an agent nonresponsible. It just denies that manipulation is the *thing* that makes an agent nonresponsible. So there must still be some relation between manipulation and the thing that makes the agent nonresponsible. But, assuming that manipulation does not make it the case that Beth is not morally responsible, why should we accept that manipulation is evidence of nonresponsibility? I contend that manipulation seems like evidence of nonresponsibility *because* it seems like it is what makes agents nonresponsible. Once we deny that manipulation makes agents nonresponsible, we need an argument that explains why it remains evidence of nonresponsibility. Without such an argument, it seems reasonable to conclude that what leads to our judgment that an agent is not morally responsible *is* what makes her not morally responsible. This, I submit, is the most plausible reading of how manipulation cases work.

Just as they lack an argument for why covert control is metaphysically, but not epistemically, irrelevant to what makes a manipulated

agent nonresponsible, incompatibilists also lack an argument for why causal determination *alone* makes manipulated agents, such as Beth, not morally responsible. So far, it seems we only have stipulations that these are true. Hence, we have yet to be given reason to believe that (a) causal determination alone makes manipulated agents not morally responsible, or (b) covert control by other agents is metaphysically irrelevant to what makes manipulated agents not morally responsible.[12]

The only argument available for (a) and (b) is further cases *without* agent-manipulators.[13] While such cases provide a reason to discount the claim that covert control *by other agents* is essential to the nonresponsibility judgment, they do not discount the claim that *covert control* is essential to the nonresponsibility judgment. It seems that any case the incompatibilist presents will have to include covert control of some kind, including cases that involve intentionless forces (see endnote 6), if she wishes it to elicit the nonresponsibility judgment. So even extra cases, it seems, do not show us that *causal determination* is the responsibility-undermining factor in a manipulation case.[14] Given this, it seems that the no generalization objection still goes through; this objection claims that we cannot generalize our judgment about a manipulated agent to a merely determined one because the responsibility-undermining factor of a manipulation case is not present in a mere determination case. Incompatibilists have yet to provide us with a convincing argument that this generalization can, in fact, be made.

So manipulation arguments do fail. But this doesn't save compatibilism, as I shall now argue.

4. Manipulation Cases

To even *prima facie* ground a manipulation argument against compatibilism, a manipulation case must be a counterexample to *all plausible current* compatibilist conditions on moral responsibility. If a manipulation case is offered that is *not* a counterexample to some compatibilist condition on moral responsibility (call it "condition *X*"), then compatibilists have an easy response: they can agree the agent is not morally responsible because she hasn't satisfied condition *X*—that is, they can provide a soft-line reply. Of course, incompatibilists have generally responded to this strategy by simply modifying their cases so that the featured manipulated agent satisfies whatever extra conditions the compatibilist can come up with (cf. McKenna 2008, p. 143). Once a case that appears to be a counterexample to all plausible current compatibilist conditions on moral responsibility is on the table, incompatibilists then posit the no difference claim (M2) and attempt to generalize the nonresponsibility judgment about a manipulated agent to a merely determined one. And it is here that the no generalization objection rears its head: we cannot successfully generalize the nonresponsibility judgment because there are *possible* further compatibilist conditions that have not been satisfied, and hence manipulation arguments fail.

But even if manipulation arguments fail, compatibilists still have a pretty major problem with manipulation *cases*. Certain of these cases, after all, seem to be counterexamples to all plausible current compatibilist conditions on moral responsibility. I take it that when compatibilists offer a set of conditions, they attempt to provide conceptually necessary and sufficient conditions on moral responsibility. If there is an apparent counterexample (or counterexamples) to all plausible current compatibilist conditions on moral responsibility, then we have three options: (1) reject compatibilism, (2) devise new compatibilist conditions—that is, defend a soft-line reply, or (3) argue that the manipulation case (or cases) is not in fact a counterexample to all plausible current compatibilist conditions on moral responsibility—that is, defend

a hard-line reply. If compatibilists lack a hard- or soft-line reply, then compatibilism is doomed.

The case we've discussed so far—Brainwashed Beth—doesn't seem to even *prima facie* support an argument against compatibilism, however. This case, as noted, is a counterexample to the leading *nonhistorical* compatibilist conditions on moral responsibility. Mele (1995; 2006), in fact, uses this sort of case to motivate his *historical* compatibilist condition on moral responsibility (though Mele claims to be agnostic between compatibilism and incompatibilism).[15] On Mele's view, an agent is not morally responsible if her recent history includes her reflective control over her mental life being bypassed in a particular way. In short, it rules that brainwashed agents like Beth are not morally responsible in line with many people's intuitions.

But we might simply invoke a case in which Mele's (and any other) historical condition is also satisfied, and we can use one of Mele's cases:

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

Is Ernie morally responsible for his *A*-ing? It seems to many that he is not morally responsible. Why? I contend that it is because Diana *covertly controls* Ernie.

Neal Tognazzini (2014) challenges the claim that covert control (or, as he puts it, manipulation) is the responsibility-undermining factor in cases such as Designed Ernie. Before continuing, it will be worth considering his challenge. He argues, instead, that it is *sourcehood* considerations that are the responsibility-undermining factor in a manipulation case. As he sees it, our judgments are sensitive to the fact that agents like Ernie are not the sources of their actions in some sense; rather, it is Diana, Ernie's designer, who is the source of his actions. This, then, highlights the incompatibilist's worry with causal determinism: if it's true, then, like Ernie, we will not be the sources of our actions even if we haven't been designed. But I think it's straightforward to see that sourcehood (or, rather, lack of sourcehood) is not the responsibility-undermining factor in a manipulation case.

Suppose that instead of Diana designing Ernie's zygote *in conjunction* with her knowledge of the past and the laws of nature, an ignorant designer designed a zygote qualitatively identical to Ernie's. The ignorant designer did so without intending that the zygote would develop in a particular way and was completely ignorant of what sort of agent would result from this zygote and what sorts of actions the resulting agent would perform. The ignorant designer still had some intentions, though: she intended that the zygote have a particular structure; perhaps she designs zygotes because she appreciates the aesthetics of zygotes. Let's call the agent that results from this zygote "Lenny." Suppose, like Ernie, that thirty years in the future, Lenny *A*s, leading to event *E*. Lenny, like Ernie, satisfies all the leading compatibilist conditions on moral responsibility. Is Lenny morally responsible for *A*-ing? At least from the perspective of an agnostic (and a compatibilist, for that matter), I contend there's not much pull toward the claim that he's not, because we have no clear reason to think that Lenny is not morally responsible. Given that we have no clear reason to think he's not morally responsible, we should hold that he is morally responsible.

If Tognazzini is correct and Ernie seems not morally responsible because he is not the source of his actions, then we ought to judge that other agents who seem to not be the source of their actions are also not morally responsible. Given that Lenny seems morally responsible, it seems that Tognazzini is committed to saying that he *is* the source of his actions. But if Erne is not the source of his actions, then neither is Lenny. After all, both have been designed by other agents; it is clear that Lenny's actions have a causal source in an another agent's intentional activity. So Lenny is not the source of his actions either. This means that sourcehood considerations are not the responsibility-undermining factor in a manipulation case. To identify the responsibility-undermining factor, we need to establish what the responsibility-relevant difference between Ernie and Lenny is. I contend that the difference is that Ernie is covertly controlled whereas Lenny is not. Design alone does not secure covert control. Covert control is secured via design *in conjunction* with a manipulator's knowledge (e.g., of the past and the laws of nature).[16]

Designed Ernie thus provides the incompatibilist with an apparent counterexample to all plausible current compatibilist conditions on moral responsibility. Manipulation cases therefore constitute objections to compatibilism in their own right. Defenders of compatibilism must still overcome these objections in order to maintain their position. Of course, manipulation arguments attempt to show that the apparent responsibility-undermining feature of manipulation cases is a feature of any determination case. In other words, there's no point in coming up with new conditions on moral responsibility because the problem is with causal determination, and not with compatibilist conditions per se. The no generalization objection suggests that compatibilists can rest easy because it shows that our intuitions do not support the claim that causal determination is the responsibility-undermining feature of manipulation cases, and therefore our intuitions do not support an argument against compatibilism. This, in effect, shows that our intuitions do not rule out *some as yet unexpressed* compatibilist conditions on moral responsibility. But given that there seem to be counterexamples to all plausible current compatibilist conditions on moral responsibility, such as Designed Ernie, incompatibilists seem to have shown that all these conditions are insufficient for moral responsibility. That is a problem for compatibilists.

5. The Control Argument

I shall now argue that counterexamples alone can undermine compatibilism. Let's think about the historicist's argument against nonhistoricism. Historicists use Brainwashed Beth (and cases like it) to argue that nonhistoricism is false. Despite satisfying the leading nonhistorical conditions on moral responsibility, it seems that post-manipulation Beth is not morally responsible. Thus, it seems that Brainwashed Beth is a counterexample to the leading nonhistorical conditions, and therefore nonhistoricism is false. Of course, if the historicist believes that Brainwashed Beth shows that nonhistoricism is false, then she clearly must think that this case supports some *argument* against nonhistoricism. This argument seems to go like this:

(H1) Brainwashed Beth is a counterexample to all plausible current nonhistoricist conditions on moral responsibility.

Therefore,

(HC) Nonhistoricism is false.

Notice that this argument contains nothing like the "no difference" premise. Thus, positions in the moral responsibility debate can be shown to be false via a manipulation-style argument *without* a "no difference" premise.

This sets the stage for an argument against compatibilism:

(D1) Designed Ernie is a counterexample to all plausible current compatibilist conditions on moral responsibility.

Therefore,

(DC) Compatibilism is false.

Call this an instance of "the control argument" (whose more general form will not specify a particular manipulation case as this instance does), as it is supported solely by a manipulation case, and I've argued that what drives the nonresponsibility judgment in manipulation cases is the fact that the agent is covertly controlled. A worry with the control argument (or any of its instances) is that it is invalid; it might seem that it does not follow that compatibilism is false just because there is a counterexample to all plausible *current* compatibilist conditions on moral responsibility. Mele, for instance, in the context of discussing manipulation arguments against compatibilism rather than motivating historicism, writes:

> Of course, even if [a manipulation case] is a counterexample to ['all alleged sets of conceptually sufficient conditions for free and morally responsible action ever proposed by compatibilists'], it does not follow that incompatibilism is true. Perhaps [a manipulation case] is not a counterexample to some superior candidate for being a set of a conceptually sufficient conditions for free action and moral responsibility that is consistent with the truth of causal determinism and that has not yet been proposed by compatibilists. (2008, p. 264)

Because a manipulation counterexample to all plausible current compatibilist conditions leaves open the possibility of some as yet unexpressed compatibilist conditions, such a manipulation counterexample apparently cannot show that compatibilism is false. An initial problem for Mele is the following. If he's correct, then he has no argument against nonhistoricism. A nonhistoricist could claim—just as Mele does with respect to apparent manipulation counterexamples to compatibilism—that a manipulation counterexample to nonhistoricism is only a counterexample to the *current* nonhistoricist conditions; it does not rule out some as yet unexpressed nonhistoricist conditions, and so the falsity of nonhistoricism has not been established.

Such a defense of nonhistoricism would be cheap, though—for surely, the burden is on the nonhistoricist to explain why apparent counterexamples to the nonhistorical conditions are not, in fact, counterexamples. This might involve positing further conditions (a soft-line reply), or it might involve arguing that manipulated agents *are* morally responsible (a hard-line reply). Unless and until the nonhistoricist shows why the apparent counterexamples are not really counterexamples, it seems acceptable to conclude that nonhistoricism is false.

Likewise, such a defense of compatibilism in response to a control argument and its associated manipulation case would be cheap. If there is an apparent counterexample to all plausible current compatibilist conditions on moral responsibility, then it seems that it is the compatibilist's burden to respond to such cases. Unless and until compatibilists respond to such cases, it seems acceptable to conclude that compatibilism is false.[17]

Indeed, there is something strange about Mele's requirement, which defenders of the no generalization objection seem to implicitly accept, that in order to show that compatibilism is false, a manipulation case must be a counterexample to *all possible* compatibilist conditions on moral responsibility. If this were a general rule about counterexamples to positions, then it seems that counterexamples would never undermine philosophical positions. After all, it seems that we will never be in an epistemic position such that we can rule out some further possible conditions on something; we're not omniscient! This is an

implausible result, and it highlights that the condition that Mele places upon counterexamples (at least in the context of discussing manipulation arguments against compatibilism rather than motivating historicism) is far too strong. That a manipulation case is a counterexample to all plausible *current* compatibilist conditions on moral responsibility constitutes a sufficient reason to reject compatibilism. This, of course, does not rule out as yet unexpressed compatibilist conditions on moral responsibility. But the mere possibility that there are such conditions, just by itself, shouldn't be taken to undermine the counterexample, and the argument it supports.

6. Conclusion

In this paper, I've argued that true threat from manipulation arguments comes from the manipulation cases used to support those arguments. These cases are apparently counterexamples to all plausible compatibilist conditions on moral responsibility. As such, they present a challenge to compatibilism in their own right. Moreover, I argued that these counterexamples can be used to support a different argument against compatibilism that I call the "control argument."

The upshot for incompatibilists is that they can present an argument against compatibilism without taking on the additional burden of manipulation arguments—namely, the notorious "no difference" premise. The upshot for compatibilists is that the no generalization objection is not sufficient to save compatibilism. What compatibilists must do, if they wish to maintain their position, is to respond to *all* manipulation cases. There is simply no getting around that.

University of Gothenburg

NOTES

Thanks to Helen Beebee and Natalie Ashton for comments and discussion on earlier versions of this paper. Thanks also to two anonymous referees for this journal.

1. Formulations of the argument along these lines are found in McKenna (2008, p. 143); Mele (2008, p. 265); Haas (2013, p. 798); King (2013, p. 67). These formulations vary, but they all move from premises like M1 and M2 to a conclusion like MC. My formulation is more detailed than these earlier formulations because I think it's worth seeing the premises that these earlier formulations suppress. I believe my formulation much more accurately represents the general structure of manipulation arguments. See Pereboom (2001; 2014) for a version of the manipulation argument – namely his Four-Case Argument.

2. King (2013, p. 66n1) notes that this case is adapted from Mele (1995, pp. 145–146).

3. For more on these conditions, see Pereboom (2001, pp. 100–110). Of course, these are only necessary conditions on being morally responsible. But, as Pereboom (2001, p. 111) points out, we can plausibly assume that a manipulated agent satisfies all the other noncontroversial conditions (at least in this context) on moral responsibility, such as epistemic conditions.

4. Hard-line replies come in various forms and have been defended by (at least) Frankfurt (2002, p. 28); McKenna (2008; 2014); Talbert (2009); Khoury (2013); and Matheson (2014).

5. Soft-line replies have been defended by (at least) Fischer (2004); Mele (2005; 2006); Baker (2006); Demetriou (2010); Waller (2014); and Barnes (2015).

6. The case I propose is similar in style to the one suggested by Pereboom (2001, pp. 115–116)—namely, one that involves an intentionless machine that comes spontaneously into existence and that then "manipulates" (or covertly controls) an agent. Gunnar Björnsson (unpublished manuscript) has

also proposed another style of manipulation case without an agent-manipulator, though his case involves psychology-changing parasites. Note that, as I allude to later, while these cases undermine Lycan's claim that what drives our intuitions is the fact that an agent is covertly controlled *by other agents*, these cases do not discount the claim that what drives our intuitions is the fact an agent is *covertly controlled*, which is the claim that I endorse below. In my view, agents are "controlled" by intentionless forces in these cases. One might object that intentionless forces can't control anything because they are not agents. However, we regularly talk of intentionless forces controlling things—for example, various processes, such as our heart rate or breathing in our bodies, are "controlled" by intentionless forces. If some continue to object to this use of "control," then they can substitute "power" for "control." It seems uncontroversial to me that an intentionless force can have power over an agent, and it is being under the power (or control) of something else that I think drives intuitions in these cases.

7. I characterize Schlosser's reply as a form of the no generalization objection because while he suggests there *might be* compatibilist conditions that can rule that designed agents are not whereas merely determined agents may be morally responsible, he never actually posits those conditions. Hence, the success of his reply, unlike a soft-line reply, does not depend on the viability of those conditions.

8. Todd's response is focused on Kearns's (2012) version of the no generalization objection. Tognazzini (2014) develops a similar, though different, line of argument in response to King's (2013) version of this objection. I consider a point of Tognazzini's below. Schlosser (2015, p. 82) replies to Todd's response in a similar way to how I do here, though mine is somewhat more developed thanks to comments from an anonymous reviewer.

9. Todd is actually talking about the Designed Ernie case, which I discuss shortly. I take it that he thinks these points apply to all covertly controlled agents, so I have applied his points to Beth in what follows.

10. The agnostic is typically appealed to in this debate by incompatibilists when compatibilists claim the debate ends in a "dialectical stalemate." Todd (2013) appeals to them in response to Fischer's (2011) claim that debate (specifically with respect to Mele's [2006] Zygote Argument, which I argue below is simply an instance of a manipulation argument) has reached such a stalemate. As does Pereboom (2008) (though he calls them "neutral inquirers") in response to McKenna's (2008) claim that the debate (specifically with respect to Pereboom's Four-Case Argument) has reached such a stalemate. See McKenna (2014) for a response to Pereboom. (Note that the term "dialectical stalemate" originates from Fischer [1994] and was used with respect to debates over the consequence argument.) Many compatibilists, however, seem to agree that Brainwashed Beth is *not* morally responsible, and so the agnostic is not typically appealed to here.

11. Thanks to an anonymous reviewer for suggesting this interpretation of Todd's position.

12. A further relevant debate has spawned here. Mele (2005) argues that the possibility of indeterministic manipulation (i.e., manipulation that has a slight chance of failing and, say, killing the agent) shows that causal determination is not what produces the judgment that a manipulated agent is not morally responsible; it is rather the fact that she has been manipulated (or, more accurately, as I've argued, that she has been covertly controlled). See Pereboom (2007) for a response to Mele, and Mele (2007) for a response to Pereboom. Note that this debate seems to assume that what produces the judgment that a manipulated agent is not morally responsible is also what makes the agent not morally responsible—for example, causal determination or covert control. It is only since Todd (2013) that incompatibilists have explicitly claimed that these things come apart.

13. As I argue in Matheson (2016, pp. 1969–1971), Pereboom's (2001) Four-Case Argument is resistant to a version of the no generalization objection (though he doesn't call it that either). Pereboom uses a *fifth* case that features no agent-manipulators to support his inference to the best explanation. Thus, the extra cases are essential to the inference to the best explanation. However, there I also argue that all

that is required is the fifth case—in other words, the inference to the best explanation is not required if one is only trying to show that compatibilism is false. Below, I suggest how the no generalization objection can be modified to avoid this move, but see also endnote 14.

14. One possible response for the incompatibilist is to argue that being causally determined is *no different* from being covertly controlled. Endorsing the move would allow incompatibilists to overcome the no generalization objection. But I won't consider this response in what follows, since I think (as I argue in the next two sections) that incompatibilists can considerably weaken their dialectical load and still have an argument against compatibilism.

15. Other historical conditions on moral responsibility have also been proposed by Fischer and Ravizza (1998) and Haji (2013).

16. Covert control can also be secured via design in conjunction with local control of the subsequent agent's (e.g., Ernie's) environment. Cf. McKenna (2000, pp. 414–415). See also Barnes (2015, pp. 560–561); he argues that Diana has "total global control" over Ernie. Also this suggests, contra Mele (2008, pp. 284–285), that manipulation and design arguments are not relevantly different; they are both instances of the same general argument form, and the only difference is the *mode* of covert control used in each. Covert control can also be secured via a more powerful intervention, such as that used by the manipulators in Pereboom's (2001) Case 1.

17. See Matheson (2014) for an attempt to defend nonhistoricism from such counterexamples.

REFERENCES

Baker, Lynne Rudder. 2006. "Moral Responsibility without Libertarianism," *Noûs*, vol. 40, no. 2, pp. 307–330.

Barnes, Eric. 2015. "Freedom, Creativity, and Manipulation," *Noûs*, vol. 49, no. 3, pp. 560–588.

Björnsson, Gunnar. Unpublished manuscript. "Manipulators, Parasites, and Generalization Arguments."

Demetriou, Kristin. 2010. "The Soft-Line Solution to Pereboom's Four-Case Argument," *Australasian Journal of Philosophy*, vol. 88, no. 4, pp. 595–617.

Fischer, John Martin. 1994. *The Metaphysics of Free Will* (Oxford, UK: Blackwell).

———. 2004. "Responsibility and Manipulation," *Journal of Ethics*, vol. 8, no. 2, pp. 145–177.

———. 2011. "The Zygote Argument Remixed," *Analysis*, vol. 71, no. 2, pp. 267–272.

Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility* (Cambridge, UK: Cambridge University Press).

Frankfurt, Harry. 1971. "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* vol. 68, no. 1, pp. 5–20.

———. 2002. "Reply to John Martin Fischer," in *Contours of Agency: Essays on Themes from Harry Frankfurt*, ed. Sarah Buss and Lee Overton (London: MIT Press), pp. 27–32.

Haas, Dan. 2013. "In Defense of Hard-Line Replies to the Multiple-Case Manipulation Argument," *Philosophical Studies*, vol. 163, no. 3, pp. 797–811.

Haji, Ishtiyaque. 2013. "Historicism, Non-Historicism, or a Mix?" *Journal of Ethics*, vol. 17, no. 3, pp. 185–204.

Haji, Ishtiyaque, and Stefaan Cuypers. 2001. "Libertarian Free Will and CNC Manipulation," *Dialectica*, vol. 55, no. 3, pp. 221–238.

Hume, David. 1739. *An Enquiry Concerning Human Understanding*. Ed. Peter Nidditch (Repr., Oxford, UK: Clarendon Press, 1978).

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

Kearns, Stephen. 2012. "Aborting the Zygote Argument," *Philosophical Studies*, vol. 160, no. 3, pp. 379–389.

Khoury, Andrew. 2013. "Synchronic and Diachronic Responsibility," *Philosophical Studies*, vol. 165, no. 3, pp. 735–752.
King, Matt. 2013. "The Problem with Manipulation," *Ethics*, vol. 124, no. 1, pp. 65–83.
Lycan, William. 1987. *Consciousness* (London: MIT Press, 1995).
Matheson, Benjamin. 2014. "Compatibilism and Personal Identity," *Philosophical Studies*, vol. 170, no. 2, pp. 317–334.
———. 2016. "In Defence of the Four-Case Argument," *Philosophical Studies*, vol. 173, no. 7, pp. 1963–1982.
———. Unpublished manuscript. "Responsibility and Independence."
McKenna, Michael. 2000. "Excerpts from John Martin Fischer's Discussion with Members of the Audience," *Journal of Ethics*, vol. 4, no. 4, pp. 408–417.
———. 2008. "A Hard-Line Reply to Pereboom's Four-Case Manipulation Argument," *Philosophy and Phenomenological Research*, vol. 77, no. 1, pp. 142–159.
———. 2014. "Resisting the Manipulation Argument: A Hard-Liner Takes It on the Chin," *Philosophy and Phenomenological Research*, vol. 89, no. 2, pp. 467–484.
Mele, Alfred. 1995. *Autonomous Agents: From Self-Control to Autonomy* (New York: Oxford University Press).
———. 2005. "A Critique of Pereboom's 'Four-Case Argument' for Incompatibilism," *Analysis*, vol. 65, no. 285, pp. 75–80.
———. 2006. *Free Will and Luck* (Oxford, UK: Oxford University Press).
———. 2007. "Free Will and Luck: Reply to Critics," *Philosophical Explorations*, vol. 10, no. 2, pp. 195–210.
———. 2008. "Manipulation, Compatibilism, and Moral Responsibility," *Journal of Ethics*, vol. 12, nos. 3–4, pp. 263–286.
Pereboom, Derk. 2001. *Living without Free Will* (Cambridge, UK: Cambridge University Press).
———. 2005. "Defending Hard Incompatibilism," *Midwest Studies in Philosophy*, vol. 29, no. 1, pp. 228–247.
———. 2007. "On Alfred Mele's *Free Will and Luck*," *Philosophical Explorations*, vol. 10, no. 2, pp. 163–172.
———. 2008. "A Hard-Line Reply to the Multiple-Case Manipulation Argument," *Philosophy and Phenomenological Research*, vol. 77, no. 1, pp. 160–170.
Schlosser, Markus. 2015. "Manipulation and the Zygote Argument: Another Reply," *Journal of Ethics*, vol. 19, no. 1, pp. 73–84.
Talbert, Matt. 2009. "Implanted Desires, Self-Formation, and Blame," *Journal of Ethics and Social Philosophy*, vol. 3, no. 2, pp. 1–18.
Taylor, Richard. 1974. *Metaphysics* (Englewood Cliffs, NJ: Prentice-Hall).
Todd, Patrick. 2013. "Defending (a Modified Version of) the Zygote Argument," *Philosophical Studies*, vol. 164, no. 1, pp. 189–203.
Tognazzini, Neal. 2014. "The Structure of a Manipulation Argument," *Ethics*, vol. 124, no. 2, pp. 358–369.
Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments* (Cambridge, MA: Harvard University Press).
Waller, Robin Repko. 2014. "The Threat of Effective Intentions to Moral Responsibility in the Zygote Argument," *Philosophia*, vol. 42, pp. 209–222.