The Problem with Manipulation*

Matt King

It is often charged that compatibilists have a problem with manipulation. There are certain cases in which victims of manipulation seem to be not responsible for what they do, despite meeting compatibilist conditions on moral responsibility. This essay argues that these arguments, as a class, fail. Their success is dependent on a particular incompatibilist assumption, one that is dialectically infelicitous in this context. My aim, however, is not to defend compatibilism but only to reject a popular argument for incompatibilism.

It is often charged that compatibilists have a problem with manipulation. There are certain cases, it is argued, where agents are the victims of manipulation and are thereby intuitively not responsible for what they do, despite meeting compatibilist conditions on moral responsibility. Thus, compatibilist conditions get the wrong result, and compatibilism is false.

This essay argues that these arguments, as a class, fail. Their success is dependent on a particular incompatibilist assumption, one that is dialectically infelicitous in this context. My aim here, however, is not to defend compatibilism but, rather, to remove one of the arrows from the incompatibilists' argumentative quiver. If manipulation arguments depend on a separate assumption, one that is already found in a large class of incompatibilist arguments, then our attention should be focused on these arguments and manipulation arguments best avoided altogether.

I. MANIPULATED BETH

My thesis is that a certain type of argument, typically pressed against compatibilists by incompatibilists, is unsuccessful. The arguments I have in mind are called *manipulation arguments,* and they concern agents who

* My thanks to Robert Kane, Michael McKenna, and Matthew Talbert for helpful comments on previous drafts. Thanks especially to Sarah Buss and the associate editors of *Ethics,* as well as two anonymous reviewers, for their comments, criticisms, and suggestions.

Ethics 124 (October 2013): 1–19

are victims of intense psychological manipulation, often involving the implantation of previously unheld attitudes, who then act on these newly acquired (but well-integrated) psychological states. Here is a version of such a case from Mele:

> 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.[1]

When Ann works diligently on a new article or volunteers for some committee service, she is plausibly praiseworthy for doing so. Not true for Beth; her newfound values are the product of manipulation, which undermines her responsibility both for them and for action that issues from them (at least action produced soon after).[2] The intuitive thought is supposed to be that manipulation undermines responsibility.

Cases such as Beth's are then used to motivate a particular sort of argument. Here's how manipulation arguments typically go:[3]

[1] Adapted from Alfred Mele, *Autonomous Agents: From Self-Control to Autonomy* (New York: Oxford University Press, 1995), 145–46. Here we should add that Beth's new values are "practically unsheddable," which means that giving up those values is not a genuine psychological option for her (172). That her values are well integrated means that she is a psychologically coherent agent whose values are not in obvious conflict (at least, anymore than is allowable in a typical, and by hypothesis, responsible agent).

[2] Whatever processes by which we can come to take responsibility for our values, the adoption of which may not have been under our direct control, may be able to render Beth responsible for her newly acquired values after living with them for a while (and engaging with and critically reflecting on them). See also Matthew Talbert, "Implanted Desires, Self-Formation, and Blame," *Journal of Ethics and Social Philosophy* 3 (2009): 1–18.

[3] For relevant examples, see Ishtiyaque Haji, *Incompatibilism's Allure: Principal Arguments for Incompatibilism* (Peterborough, ON: Broadview, 2008), 23; Michael McKenna, "Responsibility and Globally Manipulated Agents," *Philosophical Topics* 32 (2004): 169–92, 169–70; Alfred Mele, "Manipulation, Compatibilism, and Moral Responsibility," *Journal of Ethics* 12 (2008):

(M1) The manipulated agent is not morally responsible for acting on his implanted psychological states, despite satisfying compatibilist conditions on moral responsibility.

(M2) There is no difference between the manipulated agent's action and the actions of ordinary agents in a deterministic universe.[4]

(MC) Thus, compatibilist conditions on moral responsibility are insufficient. In short, compatibilism is false.[5]

The working assumption in the literature is that these sorts of cases are at least a prima facie problem for compatibilist theories of responsibility, since it is plain that such manipulated agents could satisfy the usual compatibilist conditions on responsibility but seem to be nonetheless not responsible. Beth is reasons-responsive, she identifies and endorses the values on which she acts, and her actions are expressive of her values and commitments.[6] So such theories are committed to holding that Beth is responsible. Ann/Beth is thus supposed to be a counterexample to such theories: either they've gotten the wrong result or they've got the wrong conditions on responsibility. Relatedly, compatibilist replies have typically constituted defenses of their views against manipulation cases.[7]

By contrast, incompatibilists are thought to have no trouble with Ann/Beth. The natural thought is that, given the way in which manipulation cases are constructed, it is plain that agents like Beth do not meet incompatibilist conditions on responsibility. According to incompatibilists, if determinism were true, the springs of our actions would be the result of the state of the past, even the distant past, and the interaction of

263–86, 265; Derk Pereboom, *Living without Free Will* (New York: Cambridge University Press, 2001), 110–17.

[4] It might be tempting to interpret this "no-difference principle" as claiming that there is no significant difference between manipulation and determinism. But this can't be right. Manipulation involves direct intervention by other agents, affecting the manipulated agent's psychology; while determinism is a general thesis about the causal etiology of every event. We should instead interpret the premise as claiming that there is no difference between the manipulated agent's action and the actions of determined agents, sans manipulation, consistent with the target theory. For related discussion, see Stephen Kearns, "Aborting the Zygote Argument," *Philosophical Studies* 160 (2012): 379–89.

[5] As given, this argument isn't formally valid. But the further premises required aren't relevant to my discussion here.

[6] See John Martin Fischer and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility* (New York: Cambridge University Press, 1998); Harry Frankfurt, "Free Will and the Concept of a Person," *Journal of Philosophy* 68 (1971): 5–20; R. Jay Wallace, *Responsibility and the Moral Sentiments* (Cambridge, MA: Harvard University Press, 1994); Gary Watson, "Free Agency," *Journal of Philosophy* 72 (1975): 205–20.

[7] For example, John Martin Fischer, "Responsibility, History, and Manipulation," *Journal of Ethics* 8 (2004): 145–77; McKenna, "Globally Manipulated Agents"; Talbert, "Implanted Desires."

causal forces as described by the laws of the universe. Those with incompatibilist intuitions are apt to see the result as divorcing individuals from exerting any real control over what they do. In short, we lose free will and, with it, moral responsibility.

If that is an adequate (albeit admittedly skeletal) characterization of the topography, it is no wonder that manipulation cases have appeared to be weapons for incompatibilist arguments. Beth isn't responsible, for her values are implanted without any contribution from her, just as our desires and values would be in a determined universe. Thus, for any incompatibilist account which asserts conditions on responsibility that cannot be satisfied in such a deterministic world, such an account would seemingly reach the same conclusion for Beth. Those accounts get the right verdict, so the thought goes, and manipulation cases pose a problem for compatibilists alone.

But this working view of the dialectic is mistaken. I argue that incompatibilists need to defend the truth of M1. This is for two reasons. First, compatibilists can sensibly deny the premise. Thus, an initial defense of M1 is needed to put the requisite pressure on compatibilism. More importantly, however, the second reason is that, without a defense of M1 which explains why manipulation undermines responsibility, a parallel manipulation argument can be generated for incompatibilist views. I show that this argument targets a subset of incompatibilist theories, using Kane's libertarian view as a model, suggesting that such theories are prone to the same prima facie problem compatibilists supposedly face. Thus, if M1 is left undefended, compatibilism and incompatibilism are in the same dialectical position. Or so I argue in Section II.

Defending M1, however, proves to be a difficult task. One way to do it would be to show that some condition unique to manipulation undermines the manipulated agent's responsibility. Such a proposal is unpromising, however, for any feature unique to cases of manipulation won't extend to ordinary cases of action, thus rendering M2 of the schema false. Section III illustrates this brief argument.

The remaining incompatibilist option is to claim that manipulation undermines responsibility via some feature of the cases that would generalize were determinism true. This would help ensure that M2 of the schema is true. I argue that this option is no less problematic, however, because any such feature would ensure that M2 is true only by trivializing its truth. In Section IV, I argue that the available features an incompatibilist is likely to appeal to, that manipulation determines the agent's action or that it denies the agent sourcehood over her action, are both misguided. The resulting conclusion is that the best case incompatibilists can make for manipulation arguments depends on contentious sourcehood conditions for incompatibilism, conditions that are dialectically infelicitous in this context.

One might object that incompatibilists don't need to defend the intuition that manipulated agents aren't responsible, even if that intuition is grounded only on sourcehood worries. Instead, one might think that manipulation arguments just highlight an underappreciated cost of compatibilism, one which an agnostic to the debate about free will might find compelling. I argue in Section V that this is not the case. On a plausible account of the dialectic between compatibilists and incompatibilists, the sourcehood intuition is no less contentious than the intuition supporting M1.

If all this is right, then incompatibilists cannot support M1 without invoking a contentious condition on responsibility, and without such support they are in the same boat as compatibilists. Thus, rather than privilege incompatibilism, manipulation arguments are dialectical stalemates.

II. DEFENDING THE INTUITION

Manipulation arguments begin with the claim that the manipulated agent is not responsible for acting. Beth, unlike Ann, cannot take credit for her industrious philosophizing. This claim is captured in the first premise of the schema:

> (M1) The manipulated agent is not morally responsible for acting on her implanted psychological states, despite satisfying compatibilist conditions on moral responsibility.

A potential compatibilist reply to a manipulation argument would be to simply reject M1.[8] Since intuitions may differ on these cases, M1 needs to be defended. One might insist that denying M1 illustrates a serious cost for compatibilism, holding that Beth is responsible despite severe neuronal interference. But it isn't clear that this is such a high cost to pay. First, the intuition here regards a significantly fringe case of action, involving a mechanism that is, at present, technologically impossible. One might reasonably think that our intuitions in such cases are defeasible. Second, the cost to be incurred here must be weighed against the benefits. If a compatibilist theory is able to otherwise secure our being morally responsible in a deterministic universe, this result may well be worth the cost of committing us to judging agents like Beth responsible as well.[9]

An additional reason for defending M1 is more striking. If the incompatibilist doesn't defend M1, and leaves us without an explanation for why manipulation undermines responsibility apart from the brute-

[8] As some compatibilists do. See Michael McKenna, "A Hard-Line Reply to Pereboom's 4-Case Manipulation Argument," *Philosophy and Phenomenological Research* 77 (2008): 142–59.

[9] For those that remain unconvinced, I return to the discussion of intuitions and dialectical burdens in Sec. V. For now, it suffices that this response by the compatibilist at least provisionally calls for a defense of M1 from the proponent of the manipulation argument.

ness of the intuition, then a parallel manipulation argument can be generated against incompatibilism:

(P1) The manipulated agent is not morally responsible for acting on her implanted psychological states, despite satisfying incompatibilist conditions on moral responsibility.
(P2) There is no difference between the manipulated agent's action and the actions of ordinary agents in an indeterministic universe.
(PC) Thus, incompatibilist conditions on moral responsibility are insufficient. In short, incompatibilism is false.

One might be inclined to think that this argument is insignificant, as there would be no conditions under which P1 is true. In other words, incompatibilist conditions on being morally responsible are themselves incompatible with manipulation. But this isn't necessarily true.

To see why, consider one of the principal motivations for incompatibilism. There is a thought that, if determinism is true, then when an agent acts, it is false that she could do otherwise than she does. But, so the thought goes, an agent is morally responsible for doing something only if she could have done otherwise. Thus, moral responsibility requires the ability to do otherwise, an ability made impossible if determinism were true. Some incompatibilists are motivated, at least in part, by the thought that responsibility requires genuine alternative possibilities.[10] And views which incorporate indeterminism into the conditions on responsibility to satisfy this requirement qualify as incompatibilist as a result.

Such incompatibilist accounts, however, are subject to the parallel manipulation argument above. If this is right, then a view's being incompatibilist does not mean it is better equipped to handle manipulation. As a representative instance, consider Kane's influential event-causal libertarianism. He argues that free will lies in what he calls "self-forming actions." These are actions which require the agent to consider competing self-conceptions of herself. His example is illustrative: "Consider a woman on the way to an important sales meeting who witnesses a mugging in an alley on the way. She must choose between stopping and calling for help or pressing on to avoid missing a sale crucial to her career. She is torn by conflicting motives, between self-interest and her moral values, so she must

[10] There is an enormous literature on this "alternative possibilities" condition. Two classical sources are Peter van Inwagen, "The Incompatibility of Free Will and Determinism," *Philosophical Studies* 27 (1975): 185–99, in defense of the condition, and Harry Frankfurt, "Alternate Possibilities and Moral Responsibility," *Journal of Philosophy* 66 (1969): 829–39, against. I do not take any stand here regarding whether the condition is necessary for moral responsibility, nor how to best interpret the condition. All that is required here is that there is a condition of alternate possibilities that many have taken to be necessary for moral responsibility, the satisfaction of which is incompatible with the truth of determinism.

make an 'effort of will.'"[11] Kane, being a libertarian, holds that what the businesswoman will decide is undetermined by the past and the laws of nature, though each has contributed to a causal story about the values she in fact has. Rather, it is her effort of will which will determine, albeit indeterministically, what sort of person she will be: one who helps herself or one who helps others. In Kane's view, whatever the businesswoman decides will be caused by her competing values (indeed, by the competition of her values), and so, though indeterministically caused, it will nevertheless be determined by those values. So up until the moment she decides, she might choose either option, and thus she retains fairly robust alternate possibilities. Each course of action is a real possibility for her. However, despite this view being legitimately incompatibilist, manipulation cases pose a prima facie problem.

To see this, suppose that we have two businesswomen, Ann and Beth. Ann is driven by her career ambitions but still takes seriously the needs of others. When she witnesses the mugging, she must choose between her conflicting motives of self-interest and self-sacrifice. Beth the businesswoman, in contrast, has never been particularly driven by professional ambition, though she always puts in a good day's work, nor does she go out of her way to help others, though she refrains (as much as possible) from directly harming people. Now suppose that instead of making Beth more industrious, the "brainwashing" implants Ann's competing values in Beth. Now when Beth witnesses the mugging, she experiences the same conflict of values that Ann does, although Beth's conflicting values are a direct result of the manipulation.

One may be inclined to think that in such a scenario all that effort of will is beside the point. Whatever Beth decides will ultimately be the product, not of her struggle with two of her competing values, but of her struggle with two implanted values. That is, her decision will also be the result of manipulation it seems. It is important to note here that, while she may suffer a conflict of will, just as Ann does, the conflict is between only implanted values. In this instance, all that she is deciding between is a result of the manipulation. She may very well have perfectly genuine alternatives, in the sense that it is undetermined which she will choose to do. Nevertheless, this indeterministic strand of incompatibilism does nothing to weaken the problem posed by Beth's manipulation.[12] While

[11] Adapted from Robert Kane, *The Significance of Free Will* (New York: Oxford University Press, 1996), 126.

[12] For a similar observation about Pereboom's four-case argument, see Alfred Mele, "A Critique of Pereboom's 'Four-Case Argument' for Incompatibilism," *Analysis* 65 (2005): 75–81. There he notes that the best explanation of the nonresponsibility of the manipulated agents is not that they lack alternative possibilities, for even manipulation that has only an indeterministic chance of success (but does succeed) would intuitively undermine responsibility as well.

Beth's struggle may be her struggle, the entirety of what she is struggling with has been engineered in. It would seem little consolation to explain to Beth after the fact that the decision was indeed hers, since it could only be reached via an indeterministic process. She could readily (and rightly) complain that she struggled with Ann's values, not her own.[13]

One with incompatibilist sympathies might at this point think that Beth's complaint is unfounded. He might remind her that all agents have values that are not yet theirs. Whether through inculcation, unreflective habit formation, or similar mechanisms, agents have suites of motives and values that they have not come to "own" yet. But this is just the point upon which the parallel manipulation argument plays. If there is no difference between these engineered values and ordinarily acquired ones, then, should manipulation undermine responsibility brutely, it would generalize to all cases of ordinary action (even in an indeterministic universe).

My suggestion here is that if one is inclined to think the original manipulated Beth not responsible, then one ought to think the same thing about this Beth the businesswoman. Whatever worry is posed by manipulation cases isn't isolated to compatibilist theories; even incompatibilist theories can face similar problems. If engineering in values suggests undermined responsibility by itself, then the important observation here is that one can be so manipulated and still have robust alternative possibilities. Those who think manipulation cases serve as brute counterexamples to compatibilist theories of responsibility should think the same thing about at least certain incompatibilist theories (those that rely on mere indeterminism).

This albeit limited result is nonetheless notable, for it shows that incompatibilism in general isn't itself immune from manipulation worries. Indeed, a notably incompatibilist theory like Kane's faces a problem analogous to that for the compatibilist. But it is also worth noting here that Kane's event-causal libertarianism isn't the only version of incompatibilism seemingly threatened (at least prima facie) by manipulation cases. Even agent-causal incompatibilist theories seem to face the same worry.[14] Given the example of Beth the businesswoman, there's no reason to think she couldn't have agent-caused the actions that follow her

[13] Of course, Beth in the case does not realize she has been manipulated and so could not actually voice this complaint. Nevertheless, someone with knowledge of the situation could reasonably voice it on her behalf.

[14] Ned Markosian, "A Compatibilist Version of the Theory of Agent Causation," *Pacific Philosophical Quarterly* 80 (1999): 257–77, defends an agent causal view that is nonetheless compatibilist. So the fact that a view adopts an agent causal condition doesn't thereby show that it must be incompatibilist. Nevertheless, most agent causal views are incompatibilist and would face this problem, too.

manipulation.[15] So, again, if one thinks Beth is originally not responsible for acting postmanipulation, the fact that when she acts she agent-causes herself to act would seemingly be of little importance, if manipulation undermines responsibility brutely.

The main lesson here is that a theory's being incompatibilistic does not itself make it immune to manipulation worries.[16] Even incompatibilists need some special resource to respond to the intuitive force of the manipulation case, or else they must concede that Beth is indeed responsible for her decision. What's particularly notable about this result is that this is exactly the dialectical position most compatibilists take themselves to be in.[17] When faced with a manipulation case, one can either insist that the manipulated agent is responsible, or else show why she doesn't satisfy the theory's conditions on responsibility after all. Both are ways of defeating a counterexample. But if the incompatibilist is in the same boat as the compatibilist, this would again suggest that incompatibilists face the same basic problem.

This prima facie problem for incompatibilists and the available compatibilist denial of M1 strongly suggest that M1 requires more of a defense than appeal to intuition, if manipulation arguments are going to pose a serious threat to compatibilism.[18] But not just any defense of M1

[15] That is, no reason all things being equal. We should construe Beth the businesswoman's case so that she acts in a world as described by the target theory. So, if the conditions of the target theory require agent causation, then agent causation should be possible in Beth's world.

[16] This point has also been noticed by Roger Clarke, "How to Manipulate an Incompatibilistically Free Agent," *American Philosophical Quarterly* 49 (2012): 139–49. But whereas his argument depends on a novel metaphysical case of his invention, mine relies on a familiar manipulation case used by compatibilism's opponents. Moreover, we use these observations to different purposes, relying on very different arguments.

[17] For a nice discussion of this dialectical position, see Haji, *Incompatibilism's Allure*, chap. 8; Ishtiyaque Haji and Stefaan Cuypers, "Hard- and Soft-Line Responses to Pereboom's Four-Case Argument," *Acta Analytica* 21 (2006): 19–35; McKenna, "Hard-Line Reply."

[18] John Martin Fischer, "The Zygote Argument Remixed," *Analysis* 71 (2011): 267–72, makes a similar point. His strategy, however, is quite different from my own, focusing more on the intuitions driven by manipulation cases than the structure of manipulation arguments themselves. While manipulation arguments work from manipulation cases to a conclusion of incompatibilism, Fischer works from compatibilist-friendly cases of nonmanipulation to the conclusion that manipulated agents could be responsible. His aim is thus to illustrate a defensible denial of M1 of the manipulation schema (though he doesn't put it in quite these terms). A further difference between our discussions is that he presents his case as a quasi-defense of compatibilism (whereas I remain agnostic here). For a reply both to him and to Kearns, "Aborting the Zygote Argument," see Patrick Todd, "Defending (a Modified Version of the) Zygote Argument," *Philosophical Studies* (forthcoming), doi:10.1007/s11098-011-9848-5. Fischer, Kearns, and Todd are preoccupied with a different particular argument from Mele, "Manipulation, Compatibilism, and Moral Responsibility," called the Zygote Argument (which Mele denies is a case of manipulation, although this is controversial).

will suffice. The next section places an important constraint on potential incompatibilist strategies.

III. MANIPULATION CAN'T UNDERMINE UNIQUELY

There are two rough approaches one might use for defending M1 of the manipulation schema. One, we can point to a specific feature of manipulation, unique to cases of manipulation, which helps explain why manipulated agents aren't responsible. But this approach is unpromising once we reconsider the second premise of the manipulation schema:

> (M2) There is no difference between the manipulated agent's action and the actions of ordinary agents in a deterministic universe.

If whatever explains the (supposed) nonresponsibility of manipulated agents is unique to cases of manipulation, then it plausibly won't be shared by cases of ordinary action. If what explains Beth's nonresponsibility were unique to her manipulation, M2 would be rendered false. After all, whatever one thinks of the responsibility of agents performing ordinary actions in deterministic universes, they are not manipulated. Certainly, direct alteration at the neuronal level is not the same thing as being subject to deterministic laws of nature. But this result seems to generalize to all cases of manipulation. Any feature specific to manipulation would cast doubt on there being no difference between the manipulated action and ordinary action (in a deterministic universe). Thus, incompatibilists cannot defend M1 by pointing to something manipulation-specific.

This is an important constraint on manipulation arguments. It can't be some feature unique to manipulation that supports the intuition that manipulated agents aren't responsible, for that feature will be lacking in nonmanipulated, but deterministic, contexts. So a suitable defense of M1 will have to appeal to some feature of manipulation cases that can generalize.

IV. BETH AND THE SOURCEHOOD CONDITION

Since a feature unique to manipulation is a nonstarter, incompatibilists must look elsewhere. And as we saw in Section II, pointing to alternate possibilities won't do the trick either. If manipulation undermines responsibility, it looks to do so whether or not the postmanipulated agent could do otherwise (even in an incompatibilistically respectable way). The parallel argument indicates that whatever supports the judgment that manipulation undermines responsibility, it isn't that manipulation blocks access to alternate possibilities.

Nevertheless, it might seem obvious that incompatibilists can avoid the parallel manipulation argument from Section II. First, let's rehearse that argument:

(P1) The manipulated agent is not morally responsible for acting on her implanted psychological states, despite satisfying incompatibilist conditions on moral responsibility.
(P2) There is no difference between the manipulated agent's action and the actions of ordinary agents in an indeterministic universe.
(PC) Thus, incompatibilist conditions on moral responsibility are insufficient. In short, incompatibilism is false.

As I noted at the outset, manipulation cases are often thought to pose a problem for compatibilists alone. The parallel argument shows that this isn't the case. For at least some incompatibilist views, those whose incompatibilism is driven by the indeterminism of alternate possibilities, they, too, have a prima facie problem with manipulation. But to show that some incompatibilist views are subject to a manipulation worry is not to show that all incompatibilist views are equally threatened. Indeed, many incompatibilists would not agree that Beth the businesswoman (the one subject to manipulation) satisfies their conditions on responsibility. They achieve this result by requiring more than mere indeterminism in order to be responsible. They require that the agent be the "source" of her actions. Moral responsibility for some action requires control over the causes of that action (or its preceding decision), including one's disposition, goals, values, and motivations. This is a control that determinism supposedly robs us of. We cannot control the springs of our actions, we cannot be the source of our actions, if the causal history of those actions is determined by events in the distant past. Such a concern for sourcehood, then, is a common incompatibilist refrain.

A view which adopts such a condition may appeal to it in order to deny P1 of the parallel argument. Beth the businesswoman does not control the acquiring of her new values; she isn't properly their source: the manipulation is. Thus, when she helps the mugging victim, she isn't morally responsible for doing so, and this is because she isn't properly the source of her action. If this is right, then Beth the businesswoman isn't a counterexample to the theory, and the parallel manipulation argument fails.[19]

[19] It may be worth noting that it isn't obvious that Kane's view falls into this class. On his view, self-forming actions are justified in part prospectively; that is, each is a declaration by the agent as to what sort of character and values she endorses and will live by (Kane, *Significance of Free Will*, 145). When Beth stops to help the mugging victim, this is to take one step toward taking responsibility for actions motivated by the desire to help others. The

Kane, for one, is clearly concerned with an additional motivation for incompatibilism: that an agent be the source of her actions. And part of the point of his appeal to self-forming actions is that they are meant to secure this sort of control. Similarly, Derk Pereboom holds that to be morally responsible for some action an agent must have exercised control over the decision to perform that action.[20] To exercise such control would require a special ability like agent causation, a power he finds empirically implausible. Nevertheless, we might think that a view like Pereboom's can deny that Beth the businesswoman controls her helping the mugging victim because she lacks control over the source (i.e., her values) of her action. If this is right, then incompatibilist views motivated principally by a concern about controlling the sources of action, so-called source incompatibilist views, apparently have special resources to address Beth's case. Unlike compatibilist views, manipulated Beth cannot satisfy the source incompatibilist conditions on responsibility, since at least some of those conditions require that Beth have exercised control or taken responsibility for her newly acquired values, and such conditions won't even in principle be satisfiable (at least shortly after the manipulation).[21]

However, if the parallel argument fails here, it fails due to reasons independent of the manipulation. Engineering in values is just one way agents come to have motivations for their actions. They can also unreflectively acquire them in childhood, they can be inculcated through parental instruction, habitually developed, or even spontaneously arise. What source incompatibilists must require is that the mechanism of decision and action provide a way for the agent to be the source of her action despite the varied causal history of her values, dispositions, and de-

prospective nature of responsibility for self-forming actions implies that one becomes more and more responsible for choices based on a particular value the more one acts from that value, until one is fully responsible. But this feature of Kane's view implies that he might deny P1 (of the parallel argument), not because Beth the businesswoman fails to satisfy his conditions on responsibility, but because he would find her to be responsible despite the manipulation. While her responsibility would be radically diminished, like that of a small child, she would remain responsible. Such a denial of the parallel argument, however, merely reinforces my claim that he is in the same relative position with respect to manipulation arguments as the compatibilist is. For the compatibilist can (as some compatibilists have) deny M1 by claiming that Beth is in fact responsible.

[20] Pereboom, *Living without Free Will*, 4, 43; John Martin Fischer, Robert Kane, Derk Pereboom, and Manuel Vargas, *Four Views on Free Will* (Malden, MA: Blackwell, 2007), 86.

[21] Fischer and Ravizza (*Responsibility and Control*), though compatibilists, adopt a similar condition in response to manipulation worries. See also Ishtiyaque Haji, "Authentic Springs of Action and Obligation," *Journal of Ethics* 12 (2008): 239–61. The principal difference is that source incompatibilists hold that determinism makes "taking responsibility" for the sources of one's actions impossible, whereas source compatibilists disagree.

sires.[22] For Kane, this mechanism is the self-forming action. Self-forming actions will allow Beth, after her manipulation, to come to own her implanted values and come to be fully responsible for her subsequent choices. But self-forming actions are also what allow Ann, who is not manipulated, to similarly be fully responsible for her choices based on the same values and motivations, even though these may have been acquired unfreely in childhood, say. The causal history isn't relevant to how Kane's view treats an agent's psychology; no matter the source of the values, even that they were the result of "brainwashing," through the mechanism of self-forming actions Beth takes responsibility for those values and makes them (over time) genuinely her own.

So a source incompatibilist view can claim that manipulated agents like Beth aren't responsible. They have built-in conditions to rule out P1 of the parallel argument. But these built-in conditions are a result of ruling out the responsibility of agents in a determined universe (since the views are incompatibilistic). Thus, there is no real significance to be placed on the fact that Beth's values have been implanted by the neuroscientists. Kane's view is liable to treat them as being no different from, say, Ann's values before she performs any relevant self-forming actions. It is the deliberative process, the conflict of will itself, which renders the source of one's values irrelevant to one's responsibility for Kane. This is to ensure that no matter the causal history of one's values, even if they are the unchosen product of the past or random forces of nature or one's upbringing, one can still be responsible for what one does.

But if this is the right story to tell, then whether or not that causal history involves manipulation is superfluous. It follows, then, that from the perspective of Kane's libertarianism, the details of Beth's case involving manipulation aren't specially relevant to how the view handles the case. And I think this verdict generalizes to source incompatibilist views as a class. Recall Pereboom's principal condition on responsibility, which requires that the agent have control over the source of her action, a condition which he thinks is not in principle satisfiable short of agents possessing agent causation. As we saw earlier, agent causation doesn't look helpful here, for agent-causing one's action looks to be of little help if one is acting from implanted desires and values, even if these values and desires do not determine the action all by themselves. Nevertheless, the details of Beth's manipulation aren't really relevant to how Pereboom's theory ought to handle the case; he'll simply apply his principal condition and find that Beth doesn't satisfy it.

[22] This is true for any nonimpossibilist view, as there must be some possible arrangement in which individuals can meet the theory's conditions, even if they cannot do so in the actual world. I discuss impossibilists below.

Indeed, Pereboom's view resembles Galen Strawson's, which holds that moral responsibility is in principle impossible since it would require the agent to be a *causa sui*, but nothing (short of, perhaps, God) can be a *causa sui*.[23] But even here, Beth's manipulation is beside the point, as no one can satisfy Strawson's condition on moral responsibility (or so Strawson argues). But if one is antecedently committed to requiring being a self-cause in order to be responsible, Beth's case illustrates this impossibility only as well as any case of ordinary action might. So, again, from the perspective of Strawson's view, Beth's manipulation is irrelevant.[24]

Kane, by contrast, believes we are responsible for at least some of what we do. Manipulation cases as such are irrelevant to his position, however, since manipulated agents are on a par with nonmanipulated ones from the perspective of applying his conditions on moral responsibility. The fact that an agent had values engineered in won't make her nonresponsible by itself or by direct implication. If determinism is true, then manipulated agents will be as nonresponsible as their nonmanipulated counterparts, since self-forming actions will be impossible to perform. If determinism is false, and self-forming actions are thus possible, then the causal history of a particular suite of values will be irrelevant, since it is performance of indeterministic self-forming actions which generates responsibility for action regardless of the causal history of the agent's values. Similarly, for Pereboom and Strawson, though they are inclined to hold that no one is (or perhaps even could be) responsible for what they do, this result is also independent of the manipulation the victims in manipulation cases suffer. For, again, those agents fail the respective theory's conditions just as much and for the same reasons as nonmanipulated ones.

This result is what we should expect, of course, if incompatibilists must point to a feature of manipulation that could generalize to a deterministic universe. If manipulation and determinism prevent an agent from being the source of her action, then both manipulated agents and determined ones will fail the sourcehood condition on responsibility, and for the same reasons. If such a condition really is necessary for being morally responsible, then both such agents will fail to be responsible. The upshot here is that this source incompatibilist denial of P1 of the parallel ar-

[23] See Galen Strawson, "The Impossibility of Moral Responsibility," *Philosophical Studies* 75 (1994): 5–24.

[24] To be fair, Strawson's view isn't just a form of incompatibilism. He holds moral responsibility to be incompatible with determinism, but also incompatible with indeterminism. Similarly, even Pereboom concedes that the truth of determinism is of far lesser importance than whether we can appropriately author our actions by controlling their sources. Each is thus relevant here for taking a notion of sourcehood to be the primary condition on moral responsibility.

gument rests on appealing to a sourcehood condition on responsibility. The rebuttal of the parallel argument is in turn importantly tied to a defense of the original manipulation argument. Diagnosing manipulation cases as illustrating the failure of a sourcehood condition is thus the most likely incompatibilist path to defending M1 of the original manipulation schema.

Unfortunately, this strategy is also dialectically infelicitous. If sourcehood explains why manipulated agents aren't responsible, then manipulation arguments rely on either a sourcehood intuition or a sourcehood argument. This would mean that manipulation cases piggyback on sourcehood considerations, and manipulation arguments are only as effective as their sourcehood cousins. We can, in fact, illustrate this in a more explicit fashion. First, take the manipulation schema:

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

But if manipulated agents are nonresponsible only because they fail a sourcehood condition, then the manipulation schema collapses into a sourcehood argument:

(S1) If an agent is not the proper source of action *A*, then that agent is not morally responsible for *A*.
(S2) If determinism is true, no agent can be the proper source of any action.
(SC) Therefore, if determinism is true, one can't be responsible for any action.

On this proposed incompatibilist strategy, the truth of M1 depends on the truth of S1. Similarly, M2 will depend on S2. Thus, if incompatibilists defend M1 through appeal to sourcehood, then this shows manipulation arguments to be nothing more than sourcehood arguments. If M1 is only true when an undermined-by-determinism sourcehood condition is necessary, then it is a dialectically infelicitous premise. Compatibilists already reject an incompatibilist sourcehood condition. But more to the point, one cannot legitimately defend a contentious premise through appeal to a different but equally contentious premise. Moreover, as my argument has shown, the contentiousness of the two premises arises from the same source. Though manipulation cases and a sourcehood condition may be mutually supported within some incompatibilist framework, neither can be marshaled in defense of the other in service of an incompatibilist conclusion.

If all this is right, then manipulation arguments are in trouble. If the premise that manipulated agents are not responsible goes undefended, then we're left with a clash of intuitions, and incompatibilists have their own parallel prima facie problem with manipulation. If premise M1 is to be defended, that defense must not limit itself to a feature specific to manipulation, or else the subsequent generalization premise (M2) will be undermined. Defending M1 through an incompatibilist-friendly appeal to alternate possibilities leaves us with the parallel manipulation argument. Defending M1 through an appeal to sourcehood worries gives us a coherent manipulation argument, but one that is predicated on a contentious incompatibilist assumption. Vindicating manipulation arguments thus requires defending M1, but there is no unproblematic defense forthcoming.

One might object that an incompatibilist proponent of manipulation need not appeal to an incompatibilistic understanding of sourcehood in support of M1. It is enough, the objection claims, that compatibilists agree that some kind of sourcehood is important. The idea here is that so long as a compatibilist takes seriously the significance of an agent's history, the origins of her values, dispositions, and other motivations, then one has a problem with manipulation.[25] Let us grant two competing pictures of sourcehood. Their details need not concern us. On one picture, an agent can only be the proper source of her action if determinism is false. This is the incompatibilist picture. On the competing compatibilist picture, an agent can still be the proper source of her action even if determinism is true.[26] Turning to the manipulation schema, we consider Beth, with her engineered in values, acting on them for the first time. Is she the proper source of her action? As I've understood the argument, it should be the case, if her example is constructed appropriately, that she is the proper source on the compatibilist version of sourcehood, but that she is not the proper source on the incompatibilist version. Supposing all this, whether or not Beth is indeed morally responsible when manipulated depends upon settling the general question about which sourcehood condition is correct. And this is just to reiterate my claim above: that the effectiveness of manipulation arguments rests on their sourcehood cousins. Manipulation cases, and the arguments that employ them, offer no new leverage in the debate.

[25] My thanks to an anonymous referee for posing this objection.

[26] Adoption of a compatibilistic sourcehood condition will be attractive only to those compatibilists who are concerned about sourcehood. Hierarchical views, like Frankfurt's, explicitly reject an appeal to sourcehood. Such views are likely to deny M1 by accepting that Beth is in fact responsible for her action. See Frankfurt, "Concept of a Person."

It remains open to the proponent of the manipulation schema to offer some other defense of M1, one that I haven't considered. I don't have space here to do the proponent's work for him, but it is worth re-emphasizing the challenge such a task faces. Whatever defense is to be mounted for M1, it must be amenable to generalization, per M2. But the nature of such generalization means that the defense, whatever it is, will likely render the manipulation superfluous, for whatever is true for the manipulated agent will extend to ordinary action in a deterministic universe. Thus, manipulation cases will illustrate the nonresponsibility of the manipulated agent only as well as a case of ordinary action will illustrate the nonresponsibility of a determined agent. But since the responsibility of determined agents is the very heart of the dispute between compatibilists and incompatibilists, cases of ordinary action in a deterministic world are poor tools in arguments for either side. If manipulation arguments collapse as I've argued for, then they also seem best avoided.

V. INTUITIONS AND DIALECTICAL BURDENS

I have argued that manipulation arguments ultimately fail as either insufficiently supported or dialectically infelicitous. But some might resist this conclusion. One might insist that manipulation arguments serve a burden-shifting purpose. They illustrate a cost of compatibilism: that it is committed to holding Beth responsible for her actions. And this is perhaps a cost of compatibilism that would go unnoticed in the context of determined actions, so we need to look at manipulation cases to bring it out into the open. Manipulation arguments are valuable, therefore, for making compatibilism less attractive to the unbiased observer.[27] Perhaps they do no more than better illustrate what incompatibilists have said all along about determined agents, but this might yet be a worthwhile dialectical objective.

It is unclear, however, that an agnostic about the compatibility of responsibility and determinism would find compatibilism less attractive as a result of manipulation arguments. Indeed, my discussion here has been aimed precisely at such agnostics, warning them against taking manipulation arguments as compelling. Moreover, I presume no position in the debate over moral responsibility for the purposes of this essay. As it stands, should my argument be sound, it favors neither compatibilism nor incompatibilism. All I've aimed to establish is that manipulation arguments do not count in favor of incompatibilism. So at least one way of reading the paper is to convince the agnostic that these arguments do not shift the burden to or illustrate a cost of compatibilism.

[27] My thanks to an anonymous referee for pressing me on this score.

Has it been successful in this regard? Well, consider the agnostic.[28] She must be neutral with respect to the compatibility question. She's familiar with compatibilist and incompatibilist positions while favoring neither. Suppose she is presented with the manipulation schema, and an initial case of manipulation. She reads about Beth and the neuroscientists and the brainwashing. Would she be more inclined to favor incompatibilism afterward? I find such hypotheticals difficult to evaluate.[29] But I believe there are reasons for thinking that she should not be so inclined. If she's really agnostic about compatibilism and incompatibilism, then she should also be agnostic, too, about sourcehood considerations. Otherwise, she would already find a particular argument for incompatibilism compelling. And if I'm right that the best defense of M1 relies on sourcehood, then the agnostic should find these cases unhelpful. Indeed, arguably, as she should be agnostic about a sourcehood intuition, she should remain agnostic about M1 and, therefore, remain unmoved regarding the compatibility question.

The proponent of the manipulation schema might think that the agnostic will at least concur with the following conditional: 'If the manipulated agent isn't responsible, then compatibilism is false'. Perhaps the cost to be realized by manipulation arguments is that, if manipulated agents aren't responsible, then compatibilism is refuted. But this isn't really an interesting conclusion to be reached, since everything hangs on evaluating the antecedent. Without defending M1, we can resort to upholding the conditional, but all manner of equivalent potential conditionals are possible. For example: 'If determinism rules out robust control, then compatibilism is false.' The conditionals only merit our interest if their antecedents can be defended.[30]

Ultimately, I think reflection on the agnostic actually supports my argument. Imagine the agnostic looking at two arguments: one from manipulation; the other from sourcehood. My argument implies that the former depends upon the latter. This conclusion favors neither compatibil-

[28] Appealing to an idealized agnostic is Peter van Inwagen's preferred model for philosophical argumentation, which I adopt here for illustrative purposes. See Peter van Inwagen, *The Problem of Evil* (New York: Oxford University Press, 2006), for details. See also John Martin Fischer and Neal Tognazzini, "Exploring Evil and Philosophical Failure: A Critical Notice of Peter van Inwagen's *The Problem of Evil*," *Faith and Philosophy* 24 (2007): 458–74, for discussion.

[29] For one thing, it seems that the best we can do is consult our own intuitions about what the agnostic would believe. I find little reason to suppose these intuitions to be significantly more trustworthy than our intuitions regarding the manipulated agent herself.

[30] Moreover, there are plenty of compatibilist attempts to refute M2 of the schema, arguing that even if manipulated agents aren't responsible, this result doesn't generalize to compatibilism as a whole. See Fischer, "Responsibility, History, and Manipulation"; Mele, "Manipulation, Compatibilism, and Moral Responsibility"; and Alfred Mele, "Moral Responsibility and Agents' Histories," *Philosophical Studies* 142 (2009): 161–81, for examples.

ism nor incompatibilism. Rather, it suggests that manipulation arguments are not novel attacks on compatibilism but, rather, rehearsals of sourcehood considerations. I think this argument is more likely to compel an agnostic than either incompatibilists' manipulation arguments or compatibilist defenses against manipulation arguments. But, of course, I'm biased.

QUERY TO THE AUTHOR

No Query.