Philos Stud
DOI 10.1007/s11098-016-0754-8

Defeating Manipulation Arguments: Interventionist causation and compatibilist sourcehood

Oisín Deery[1] · **Eddy Nahmias**[2]

Abstract We use recent interventionist theories of causation to develop a compatibilist account of causal sourcehood, which provides a response to Manipulation Arguments for the incompatibility of free will and determinism. Our account explains the difference between manipulation and determinism, against the claim of Manipulation Arguments that there is no relevant difference. Interventionism allows us to see that causal determinism does not mean that variables outside of the agent causally explain her actions better than variables within the agent, whereas the causal source of a manipulated agent's actions instead lies outside of the agent in the intentions of the manipulator. As a result, determined agents can have free will and be morally responsible in a way that manipulated agents cannot, contrary to what Manipulation Arguments conclude. In this way, our account demonstrates not only how Manipulation Arguments fail but also how compatibilism can be strengthened by means of a plausible account of causal sourcehood.

Keywords Free will · Moral responsibility · Determinism · Manipulation Argument · Zygote Argument · Interventionism · Causation

Oisín Deery and Eddy Nahmias: Authorship is equal.

✉ Oisín Deery
oisin@oisindeery.com

[1] Department of Philosophy, University of Arizona, 1145 E. South Campus Drive, Tucson, AZ 85721, USA

[2] Department of Philosophy & Neuroscience Institute, Georgia State University, 25 Park Place, Suite 1600, Atlanta, GA 30303, USA

Published online: 08 August 2016

Causation is a morass in which I for one refuse to set foot. Or not unless I am pushed.
—Peter van Inwagen (1983: 65)

… man is the source and begetter of his actions as a father is of his children… If we cannot trace back our actions to starting points other than those within ourselves, all actions in which the initiative lies in ourselves are in our power.
—Aristotle (Nicomachean Ethics III: 5)

1 Introduction

Free will is often defined in terms of the *causal* efficacy of agents and their deliberations, and debates about free will often focus on its relation to *causal* determinism. Accordingly, it should not take much to push philosophers debating free will into the "morass" of causation. Yet most have followed van Inwagen's lead by setting aside questions about causation or by aiming to remain neutral on them.[1] Given the complexity of debates about causation, this strategy is understandable. Yet given the centrality of the notion of causation to free will, the strategy is ill-advised.

We argue that recent *interventionist* theories of causation provide a response to Manipulation Arguments for the incompatibility of free will and determinism, by developing a compatibilist account of *causal sourcehood*. Our account explains the difference between manipulation and determinism, against the claim of Manipulation Arguments that there is no relevant difference. Even if causal determinism is true, we argue, an agent can be the causal source of her actions, since often no variable beyond the agent's control will have a stronger causal–explanatory relationship with her actions than relevant variables within her control. On the other hand, the causal source of a manipulated agent's actions lies beyond the agent's control in the intentions of the manipulator. As a result, determined agents can be free and responsible, contrary to what Manipulation Arguments conclude, whereas manipulated agents have, at best, reduced freedom and responsibility. In this way, we demonstrate not only how Manipulation Arguments fail, but also how compatibilism can be strengthened by providing a plausible account of causal sourcehood.

2 The Manipulation Argument

Our account answers the most significant contemporary arguments for incompatibilism about free will and determinism, Manipulation Arguments. In deploying this style of argument, *source incompatibilists* have shifted attention away from the problem that determinism appears to pose for an agent's having alternative possibilities and instead toward the problem that it poses for an agent's being the causal source of her actions. Source incompatibilists thereby ignore van Inwagen's

[1] Exceptions include some agent-causal libertarians (e.g., O'Connor 1995), and more recently compatibilists such as Vihvelin (2013), Sartorio (e.g., 2013), Ismael (2013) and Roskies (2012).

injunction to avoid the morass of causation, since they suggest that the problem posed by determinism is that when an agent makes a decision, "he is causally determined by factors beyond his control to decide as he does" (Pereboom 2014: 79).

We agree with source incompatibilists that causal sourcehood is relevant to free will and moral responsibility. Yet we flip the incompatibilist's argument on its head in order to illuminate a *compatibilist* account of causal sourcehood.

To get a sense for how our argument will work, imagine a pair of cases (drawn from Mele 2013, 2006). First, imagine Danny. One evening in 1986, Danny's parents made love, hoping to conceive a child. They got lucky. A zygote was formed (at time t_1), and nine months later Danny was born. Thirty years later, Danny is walking down a deserted street and he finds a wallet with the owner's ID in it and $500. Danny takes himself to have good reasons for keeping the money, but also for returning the wallet. He deliberates for a while, and in the end he decides to keep the money, and he does so (at time t_{30}). Assume that this occurs in a *deterministic* universe—that is, a universe in which, for each event E, the laws of nature and some set of events that occurred prior to E are such that these events cause E to occur with probability 1.[2] If determinism is true, then some set of events prior to Danny's act of stealing the wallet at t_{30} are (together with the laws) such that they cause his deliberating and acting in that way, at that time, with probability 1.

Now imagine a different case, in which a powerful Goddess, Diana, has the power to know what will happen in the future and to act in ways that ensure that specific events occur in the distant future. Diana has these abilities in part because she exists in a deterministic universe and is able to get enough information about events occurring in it (e.g., at t_1) to deduce exactly what she needs to do at that time to ensure that a particular event occurs 30 years later. In this case, Diana assembles atoms in a specific way at t_1 so as to create a zygote that develops into a child, grows up, finds a wallet 30 years later, and at t_{30} decides to keep the money it contains. For some reason, Diana wants to ensure that this event occurs at t_{30}, and she possesses the power to alter events at t_1 precisely so that she ensures that it does occur.[3]

As it turns out, the life of this intentionally created person (whom we will call Manny since he is *Man*ipulated) follows the exact same course as the life of deterministic Danny (as described above). Manny is no different from Danny when it comes to his abilities to consider options, weigh reasons, and make decisions about whether to steal the money. The Manipulation Argument maintains that Manny lacks free will in making his decision, and so he is not morally responsible for it. The argument then maintains that there is no difference between Manny and Danny that is relevant to free will and moral responsibility, and as a result we should recognize that Danny is not free or responsible either. In both cases, the agents are causally determined to perform actions by factors ultimately beyond their control.

[2] Determinism is sometimes defined in terms of *entailment*: a description of the universe at a time, in conjunction with the laws of nature, entails descriptions of the universe at other times (e.g., van Inwagen 1983).

[3] We will accept for the sake of argument the stipulation that Diana, as described, is nomologically possible, despite its being questionable whether she is.

Why would incompatibilists use Manipulation Arguments to advance their cause? At first glance, their doing so seems ill-advised, since there is a long history of compatibilists' pointing out that it is misleading to present determinism in terms of the past or laws controlling or compelling us, as manipulators do (e.g., Dennett 1984; Ayer 1954). There are obvious differences between real-world manipulation and determinism (or any ordinary causal history), since manipulation typically includes a powerful agent who is capable of altering her dupe's external or internal conditions to ensure that the dupe does what she wants. Often, such manipulation involves changing what the manipulated agent would otherwise want to do or bypassing the agent's rational capacities. The more powerful the manipulator, the more inevitable the outcome—one way or another, the dupe does exactly what the manipulator wants.

Manipulation Arguments attempt to strip away some of these features of real-world manipulation, so that the dupe maintains his compatibilist capacities rather than having them compromised or bypassed. Whether we consider Plum in Derk Pereboom's Four-Case version of the argument (Pereboom 2014, 2001) or Ernie in Alfred Mele's Zygote Argument (2013, 2006), the manipulated agent is presented as satisfying any combination of compatibilist conditions proposed as minimally sufficient for free will and moral responsibility. At the very least, the agent possesses the ability to reflect on and identify with his desires (Frankfurt 1971) and the capacities for recognizing and responding to reasons (Fischer and Ravizza 1998), including moral reasons (Wolf 1987), while not acting on compulsive or compelled desires (Mele 1995). From the inside, the manipulated agent feels as free as anyone else, and from the outside he behaves just like a normal agent.

Combining these features (and perhaps others offered by compatibilists), we can say that the manipulated agent's decision is proximally caused by the output of his "Compatibilist Agential Structure," or CAS (McKenna 2008)—that is, by features of the agent's psychology that compatibilists typically judge as jointly (and minimally) sufficient for free will and moral responsibility. Once the Manipulation Argument clarifies that the manipulator works *through* the dupe's CAS in this way, it becomes difficult to find a principled difference between manipulated histories and ordinary causal histories. Indeed, Pereboom aims to strip away *every* relevant difference in order to leave causal determination by factors beyond the agent's control as the only similarity that can be used to explain why the agent lacks the sort of free will required to deserve praise or blame for her action (what Pereboom calls "basic desert"). By contrast, Mele leaves it open what might explain the intuition that the manipulated agent is not free or responsible, and he remains agnostic about whether this intuition is justified. The clearest way to present the cases so that the agent acts *through* his CAS, rather than having the manipulator bypass any of the relevant capacities at, or before, the time of action, is to have the manipulator *design* the agent ahead of time so that he makes the desired decision while engaging his CAS (as in Pereboom's Case 2). Mele's Zygote Argument offers an especially clear presentation of this element of the Manipulation Argument, and therefore we focus on it here, but our response generalizes to other versions of the argument, including Pereboom's. Here is Mele's version, using the cases that we described above:

(1) In deciding to keep the wallet at t_{30}, Manny does not have free will and is not morally responsible. (More generally, any agent who is manipulated or created in this way so that he will decide to do action *A* lacks free will and moral responsibility regarding his *A-ing*, even if he acts through his CAS.) [Call this the *NoFW Premise*.]

(2) Regarding free will and moral responsibility, there is no principled difference between Manny and Danny. (More generally, systematic manipulation or creation of this sort is no different than determinism when it comes to free will and moral responsibility.) [Call this the *NoDif Premise*.]

(3) So, Danny does not have free will or moral responsibility in deciding to keep the wallet at t_{30}. (More generally, free will and moral responsibility are incompatible with determinism.)[4]

McKenna (2008) calls compatibilist responses to the Manipulation Argument that reject the *NoFW Premise* "hard-line" responses, and those that instead reject the *NoDif Premise* "soft-line" responses (see also Kane 1996). The dominant response to the argument from both sides of the debate has been to accept the *NoDif Premise*, and subsequently to disagree about whether Manny has free will and is morally responsible. Source incompatibilists maintain that Manny lacks freedom and responsibility, generating the incompatibilist conclusion. By contrast, hard-line compatibilists argue that Manny *can* be free and responsible, since he satisfies all compatibilist conditions, despite his having been manipulated (McKenna 2008). Hard-liners see no way to avoid biting the bullet here. As Gary Watson puts it, "It is hard to see what differences there could be between the natural and purposeful forms of determination that would be relevant to freedom and control" (2000: 67). John Fischer puts the point even more forcefully:

> I think that manipulation cases are compatibilism's dirty little secret. Compatibilists don't like to admit that this is a problem. It is to Bob Kane's and other incompatibilists' credit that they have pushed us to confront cases of covert non-constraining control. …We compatibilists have to deal with this. (2000: 390)

As compatibilists, we agree that Manny satisfies most of the requirements for free will and moral responsibility. Indeed, our account shows why Manny is a causal difference-maker regarding his action of stealing. Yet we agree with source incompatibilists that Manny lacks an important component of free will and responsibility—namely, a type of causal sourcehood for his actions. Hence, we reject the *NoDif Premise* and instead advocate "soft-line" compatibilism.[5]

[4] We use "Manipulation Argument" broadly to include Mele's Zygote Argument. Debates have raged about whose intuitions are relevant to these arguments (e.g., McKenna 2014b, 2008; Mele 2013; Pereboom 2008). We take the consensus view that our audience is well-informed readers who are undecided (or neutral) about the compatibility of determinism and free will.

[5] McKenna defends the hard-line response, yet maintains that the compatibilist "needs to hold that an agent can be the ultimate source of her will and her acts even if there is a deterministic explanation for them that traces back to factors for which she is *not* ultimately responsible" (2014a: 85). We agree that determinism does not rule out the sort of sourcehood relevant to free will. If McKenna came to agree with us that *some* forms of intentional manipulation *do* rule out this sort of sourcehood, perhaps he would agree that our soft-line response should be deployed against arguments that appeal to such forms of manipulation.

Currently, advocates of this position include Schlosser (2015), Mickelson (2015)/ Demetriou (2010), Waller (2014), Barnes (2013), Sripada (2012), Ragland (2011), and Haji and Cuypers (2006). Each of these philosophers emphasizes that there is a relevant difference between determinism and intentional manipulation or creation by another agent. Our account follows suit, yet we go further by providing a more systematic account of why this difference exists, and why it makes a difference to freedom and responsibility.[6] The difference is that intentional manipulation introduces a causal source for the decision that is beyond the agent's control, whereas determinism does not. This difference will be illuminated by the interventionist theory of causation.

3 Interventionism, causal modeling, and strength of invariance

Interventionist theories of causation are increasingly influential. In debates about the metaphysics of causation, Hitchcock (2001) has argued that interventionism deals with a number of problematic counterexamples that plague other theories of causation. Additionally, McCain (2012) has used the framework to develop a causal account of the metaphysical basing relation, while Yang (2013) maintains that arguments against the existence of composite objects fail because interventionism reveals that such objects exist. When it comes to agency, Woodward (2015) and List and Menzies (2009) defend interventionist accounts of mental causation in response to the causal exclusion problem. Finally, some philosophers have begun using interventionism to think about free will (e.g., Ismael 2013; Roskies 2012; Campbell 2010), and the theory has also proven useful in understanding people's judgments about causation and moral responsibility (e.g., Lagnado et al. 2013; Sloman 2005). While we cannot offer a comprehensive defense of interventionism here, we believe that it is a highly plausible account of causation and we will argue that it provides an attractive compatibilist response to the Manipulation Argument.

Suppose that we are trying to discover the *causal source* of an event. In doing so, we must consider which of the many prior causes that influence it has the most significant causal relation with, and provides the best causal explanation of, that event. In the case of agents performing actions, the relevant causal factors will usually be mental events occurring within the agents. Yet in order to determine whether such events are the causal sources of Danny's and Manny's actions, first we need to know whether there are earlier causes that have even *stronger* causal–explanatory relations with these actions. Interventionism yields a sophisticated way of doing precisely this, since it models *how* an event outcome depends on prior causal factors. Woodward (2003) has developed a philosophically rich framework

[6] We do not take a stand on whether Manny lacks free will *entirely* or whether he has *less* or a lesser sort than Danny. We maintain that the Manipulation Argument fails if the *NoDif Premise* is false, and, contra Todd (2011), that offering a principled difference between manipulation and determinism that is relevant to the *degree* to which an agent is free or responsible is enough to show that this premise is false. For discussions of degrees of moral responsibility, see Nelkin (2016), Khoury (2014), Capes (2013), Coates and Swenson (2013) and Tierney (2013).

of this sort, drawing on the work of Pearl (2000) and others (e.g., Menzies 2004). We appeal to this framework, which we explain below, in order to provide a novel account of causal sourcehood.

Within the interventionist framework, when we consider whether something is a cause, we ask, "What if things had been different?" and by answering this question we identify factors whose manipulation would produce changes in the outcome being explained. If this (cause) variable were altered in these given ways, then this (effect) variable would be altered in these other ways. This information is derived from a *causal model*, which is a representation that encodes counterfactual dependency relationships among variables, where the variables represent event-types in such a way that they can be set to different values (representing particular token events) by interventions.

There are two central ideas within this framework: *intervention* and *invariance*. To explain them, we will develop causal models for both Danny's and Manny's cases. Let us start with Danny.

According to interventionism, we must first select variables for our model. For instance, we need a variable representing the event of Danny's CAS having a certain output just prior to his stealing the wallet (call this variable X for now), and we need a variable representing whether Danny subsequently steals the wallet (call it Y for now). An *intervention* is an exogenous experimental manipulation of X to find out whether X causes Y. Such interventions are "surgical," in the sense that the usual causes of X, or of its taking a particular value, are ignored (for a fuller account of what counts as an intervention, see Woodward 2003: 12–17; Chap 3). In this framework, X causes Y just in case, for at least some state of the model, there is an intervention on X that would reliably change the value of Y. If so, then as Woodward puts it, X is a *direct* cause of Y (for the notion of direct causation, see Woodward 2003: 52–61).

Interventionism enables us to test the causal contributions of *any* causal variable while ignoring its prior causes. However, if our aim is instead to pick out the *causal source* of an event's occurring, we need to know how far back to trace when identifying causal contributions. For instance, we may need to identify a causal source when we want to develop effective strategies for controlling the values of an outcome variable (Cartwright 1979), and in the context of moral (or legal) responsibility practices, we typically want to identify a causal source because we want to locate the appropriate targets of our blame and praise (and potential punishment). To identify the causal source of any event's occurring, we need to identify the causal variable that has the most stable causal–explanatory relation with the effect variable. To develop such a notion of causal sourcehood, we must understand *causal invariance* (see Woodward 2003: 12–17, 69–70, 119, 183–4; Chap 6).

A causal generalization, G, relating variables X and Y, counts as causally invariant to various degrees just in case G describes how Y would change under a range of alterations to (or interventions on) the value of X. These invariance relations are stated as structural equations, which are asymmetrical in that the value of the variable on the left hand side of the "=" is determined by what appears on the right hand side:

$$Y = f(X, \ldots, Z_n)$$

Here, the Z_n (if any) are just the other causal variables—or direct causes—relevant to Y that are explicitly represented in the model, while the function, f, for a structural equation is expressed as a mathematical operation on the variables. The equations in any model encode (often multiple) counterfactuals of the following form (where X, Y, and Z represent variables, and x, y, and z represent the respective values that these variables take):

> If it were the case that the value of $X = x$,…, and $Z_n = z_n$, then it would be the case that the value of $Y = f(X,\ldots, Z_n)$, in background conditions C.

The background conditions, C, are all the direct causes of Y that are not explicitly represented in the model (due to their not explaining why Y takes one particular value rather than another; more on this below).

Thus, a causal model consists in a set of structural equations, where the equations express invariance relations obtaining between the variables in the model. A model can also be represented graphically, as depicted in Fig. 1.

Fig. 1

Here, X is depicted as a direct cause of Y, and the arrow indicates this causal–explanatory relationship.

The next step in identifying the causal source of an event's occurring, once we have established a direct causal relationship, is to test for *actual* causation (Woodward 2003: 39–40, 74–86), so as to establish what causes what in an actual chain of events. Actual causation is defined in terms of a "sequence" of causal variables within a model, or of direct causal "arrows" in its graphical representation. Thus, the expression "X's taking the value x is an actual cause of Y's taking the value y" means that there is at least one sequence of variables from X to Y (where this sequence may include just X and Y or also multiple intermediate variables) for which an intervention on X will change the value of Y, given that other direct causes of Y within the model—and that are not part of this sequence—are held fixed at their actual values. If so, then $X = x$ is indeed an actual cause of $Y = y$.

Having established actual causation, it remains a further question, we maintain, whether $X = x$ is the *causal source* of $Y = y$. Here again, we must consider invariance. An additional component is required, which is that the invariance relation between X and Y remain stable not only under interventions on X, but also under a range of alterations to the background conditions, C. This requirement enables us to identify *strength of invariance*, which is a central component of our definition of causal sourcehood.

More precisely, a causal invariance relation, R_1, that obtains between two causal variables, X and Y, is stronger than another such relation, R_2, obtaining between Y and another of its prior causal variables—for instance, W—iff:

(1) holding fixed the relevant background conditions, C, R_1 predicts the value of Y under a wider range of interventions on X than R_2 does under interventions on W; and
(2) R_1 predicts the value of Y across a wider range of relevant changes to the values of C than R_2 does.

In what follows, condition (2) will be of particular importance.[7] What this condition says is that the counterfactuals encoded in the R_1 equation—which express the invariance relation obtaining between X and the outcome variable Y—hold under a wider range of counterfactual situations than those encoded in the R_2 equation. Put simply, variations on X would cause variations in Y across a wider range of relevant background conditions. Using these two conditions, we can define the notion of causal sourcehood:

> $X = x$ is the causal source of $Y = y$ iff X bears the strongest causal invariance relation to Y among all the prior causal variables (including X) that bear such relationships to Y.

Defenders of the Manipulation Argument agree with us that an agent must be the causal source of her action in order to act freely. However, their argument tries to show that no agent is the causal source of her action in a deterministic world. As we will explain further in Sect. 5, advocates of the Manipulation Argument typically define sourcehood in a way that is ad hoc and not generalizable (e.g., Pereboom 2014: 74). By contrast, our definition provides a general way of discerning which causal variables in a system are more causally significant than others regarding an outcome event, and in a way that accommodates the idea that sourcehood, like invariance, can come in degrees.[8] The central idea is that causal variables that can result in an outcome by more than one means—that is, in response to a wider range of changes to the background conditions—bear stronger causal invariance relations to their outcome variables than variables that cannot, or than variables that only result in the outcome across a narrower range of changes to the background conditions. For Danny there is a strong causal invariance relation obtaining between the output of the deliberative activity occurring within his CAS and his subsequent act of stealing the wallet, and likewise for Manny.

As Lombrozo (2010: 309–10) puts it, the deliberative activity of intentional agents like Manny or Danny exhibits, to a strong degree, *equifinality*—it results in a particular outcome (the intended one) across many different conditions (e.g., Romeo's decision to reach Juliet will result in his doing so despite many possible

[7] See footnote 10, below, for an explanation of why this is so.

[8] We are defining causal sourcehood in terms of the causal variable that has the strongest invariance relation to the effect variable. However, because strength of invariance allows for relative comparisons between causal variables, we could also define *relative* causal sourcehood among selected causal variables (see also footnote 13). In cases in which the strength of invariance obtaining between Diana's decision and Manny's stealing is equal to that obtaining between the output of Manny's CAS and his stealing, Diana and Manny may share equal responsibility. Similarly, different causal contributions may result in assignments of different degrees of responsibility to Diana and Manny (see footnote 6). This feature of our view enables us to deal with cases in which Diana's decision is, for instance, only slightly more or less invariant as a cause of Manny's stealing than the output of Manny's own CAS is.

obstacles). The behavior of non-intentional objects, by contrast, exhibits greater *multifinality*—different conditions result in different outcomes (e.g., iron filings will not reach the magnet to which they are attracted even if a simple obstacle, such as a card, is placed between them).[9] In this way, equifinal causal variables typically bear strong causal invariance relations to their outcome variables. That is why it is correct to judge the outputs of both Danny's and Manny's deliberations as *actual* causes of their actions, ones with relatively strong invariance relations with those actions.

Of course, this consideration might tempt one toward the hard-line response to the Manipulation Argument, since in each of Danny's and Manny's cases we can carve off the variable representing the output of deliberation from its antecedent causes, with the result that the output of activity in their CASs counts (in both cases) as an actual cause of the stealing. (We will discuss this issue further in Sect. 5.)

However, we are not only seeking actual causes when we ask whether an agent acts freely, as Manipulation Arguments reveal. Many intermediate events, including non-mental events within the agent, might occur between the agent's deliberation and her subsequent action—for instance, electrical impulses in the muscles of the arm that reaches for the wallet—and these events might be modeled as actual causes of the bodily behavior. Yet while such events count as actual causes, and may be useful to consider in some explanatory contexts, they do not count as causal sources of the action, since causal sourcehood depends on strength of invariance (note, however, that it is still a *necessary* condition on an event's being the causal source of another event that it be an actual cause of that event). The output of deliberative activity within an agent's CAS typically bears a much stronger causal invariance relation with her actions than, for instance, muscle movements in an arm, because the CAS's output will cause the action across a wider range of circumstances. Given the goal of stealing the wallet, Danny and Manny will achieve this goal even if they need to move their arms in a number of different ways to do so (cf. Campbell 2010).

Yet as strong as the causal relation between Manny's CAS and his act of stealing is, there exists an even stronger causal invariance relation between Diana's decision that Manny steal and Manny's subsequent action. The relation between Diana's decision and Manny's action is such that, across a maximally wide range of changes to the background conditions, the variable representing Manny's stealing does not change in value without a change in the value of the variable representing Diana's decision, and changes in the value of the variable representing Diana's decision correspondingly change the value of the variable representing Manny's decision to steal (or that of another agent she designs to carry out her desired outcome at t_{30}).

Conversely, when we model Danny's situation, there is no variable in *his* distant past that bears such a causal invariance relation with *his* decision—certainly, there is no variable that bears a stronger invariance to his stealing than the output of Danny's own CAS does. Recall, we are considering a case in which all other compatibilist conditions are met, which we take to include that there is no external cause that determines Danny's action while bypassing his internal deliberative processes. If we sought to locate the strongest causal relation affecting Danny's

[9] The examples of Romeo and the iron filings are drawn from James (1890: 20).

action, we would focus on the activity occurring in his CAS just prior to that action, and the output of that activity, rather than on any prior causal factors. There is no variable, for instance, at t_1 whose value could be varied such that it would cause Danny to decide to steal at t_{30} across a wide range of background conditions (or whose value could be varied in systematic ways such that it would vary Danny's actions at t_{30} systematically, while *holding fixed* C). As a result, the causal source of Danny's decision lies within him, not in the distant past. The best explanatory model for Danny is the one depicted in Fig. 2, according to which the output of activity in Danny's CAS, which is represented by *Danny* and takes a particular value, *d*, is the causal source of his stealing the wallet, which is represented by *Steal* and takes the value *s*. This is because there is no variable prior to *Danny* that bears a stronger causal invariance relation to *Steal* than *Danny* does, according to our strength-of-invariance conditions. In other words, the actual output of Danny's CAS is the causal source of his stealing the wallet.

Fig. 2

By contrast, Diana's decision, as represented in Fig. 3 by the variable *DD*, which takes the value *d*, is the causal source of Manny's stealing, again represented by *Steal*. This is because *DD* bears a stronger invariance relation to *Steal* than the output of activity in Manny's own CAS does, here represented by *Manny*. Had Diana decided that she wanted Manny to *return* the wallet, then she would have created his zygote in a different way so that he would decide to return it. Furthermore, since we are assuming that Diana is able to *ensure* (30 years prior to the event in question) what Manny will do, there are no relevant changes to conditions *C* that could possibly interfere with Manny's stealing, since Diana (we are assuming, along with advocates of the Manipulation Argument) has foreseen all such possibilities.[10] As a result, *DD* bears a stronger causal invariance relation to *Steal* than any other prior causal variable does, including *Manny*.

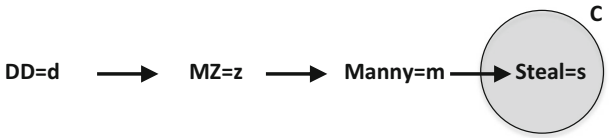

Fig. 3

[10] This is why the second of our two conditions on strength of causal invariance is important. In the case of Diana and Manny, condition (1) alone cannot explain whether *DD* or instead *Manny* is more strongly invariant regarding *Steal*. Yet in cases of "online" manipulation, where things do not work out as Diana had intended, such that she has to remain online and ready to step in again to ensure the outcome, condition (1) *always* shows that *DD* is more strongly invariant. The Manipulation Argument, however, is about "offline" manipulation—Diana makes an initial change to ensure that *Steal* takes a specific value (via *Manny*), after which she goes offline and allows events to unfold without stepping in again. As we will show, condition (2) explains why, in offline manipulation, *DD* is more strongly invariant regarding *Steal* than *Manny* is.

Diana creates Manny's zygote (event *MZ*) in such a way (*z*) that Manny's own deliberations (*Manny* = *m*) bring about exactly what she wants (*Steal* = *s*). The reason that Diana's decision (*DD*) bears the strongest causal invariance relation to *Steal* is that Diana can (we are assuming) ensure that Manny steals as she intends across the widest possible range of changes to *C*.

By contrast, if we intervene on *Manny* to test its strength of causal invariance with *Steal*, we must ignore the contribution of *DD*—that is simply what an intervention requires. Now imagine that Manny's sister, Franny, who is an astronaut on the first human expedition to Mars, sends Manny a text message as she touches down on the red planet. This counts as a relevant change to the background conditions that bear on *Steal*. Assume that Mars is at its nearest approach to Earth—i.e., three light-minutes away. Franny sends her message exactly three minutes prior to the instant when Manny bends down to steal the wallet. Seeing the message from Franny, Manny is so overcome by pride in his sister's achievement that he forgets all about the wallet and strolls away. Here, it is nomologically impossible for activity in Manny's CAS to produce an output that controls for the influence that Franny's message has on whether he steals (he cannot know about this message before it arrives).

Yet, as we saw in testing for the causal invariance that *DD* bears to *Steal*, Diana *is* able to control for the influence of an event like the arrival of Franny's message (by stipulation of the Manipulation Argument), since Diana can foresee everything that could possibly bear on Manny's action and can control for what she foresees.

Importantly, notice that the crucial difference between Manny and Danny is not intrinsic to them. Throughout their lives up to their respective decisions to steal, Manny and Danny are intrinsically identical, as are their relevant universes going back to their zygotes' creation. However, there are *extrinsic* causal facts about Manny and Danny that differ radically, and which explain the differences in causal sourcehood. Namely, Manny's universe includes a powerful manipulator who causally influences what happens in that universe.

Of course, it is not enough simply that Manny's universe *include* Diana, since Diana might exist but decide to do nothing to affect Manny. In that case, and assuming that Manny later steals, we *could* intervene on a variable representing Diana's decision in a way that would change the value of *Steal* across as wide a range of changes to *C* as in the regular case. Would Diana's deciding to do nothing thereby be the causal source of Manny's stealing (*Steal* = *s*)? No, since for any variable *X*'s taking a value, *x*, to be the causal source of another variable *Y*'s taking a particular value, *y*, $X = x$ must at least *actually* cause $Y = y$. If Diana decides to do nothing to affect Manny, then obviously her decision is not an actual cause of *Steal* = *s*, and so it is not the causal source.[11]

[11] Or imagine that Zeus, a more powerful deity than Diana, does not care about whether Manny steals the wallet. Yet he is settled on this: Either Diana decides to create Manny's zygote in a particular way (which, as it happens, will result in Manny's stealing the wallet) or Zeus will interfere to make Diana do so (i.e., Zeus is a Frankfurt-style intervener). In the actual case, Diana decides on her own; Zeus does nothing. Is Zeus the causal source of Manny's stealing? No. Zeus's deciding to do nothing is not an actual cause of *Steal* = *s*. So it is not the causal source of *Steal* = *s*, even though had Diana decided differently, it *would* have been. (Thanks to Al Mele for suggesting this case.) Similar reasoning indicates how our view offers

As we show in the next section, the difference in causal facts between Danny's and Manny's universes provides a principled soft-line response to the Manipulation Argument. To the extent that the strongest causal invariance relation obtaining between Manny's stealing and any prior causal variable traces back to Diana's intentions, we advocate such a soft-line response. Otherwise, as we will explain in Sect. 5, our account tells us precisely when, and why, we should instead opt for the hard-line response.

4 Compatibilist causal sourcehood and a principled soft-line response

We will now present our argument in a more systematic way. First, we agree with source incompatibilists that it is relevant to an agent's free will and moral responsibility for performing an action whether the action has its causal source in the agent. We also agree with source incompatibilists—and their Manipulation Arguments—that agents who are intentionally manipulated in the way that Manny is lack free will and moral responsibility, since such agents are not the causal sources of their actions. Yet whereas defenders of the Manipulation Argument claim that the reason that Manny is not free or responsible reveals that the same is true of Danny, we offer an explanation for why Manny lacks free will and responsibility that does not generalize to Danny, and hence does not support incompatibilism.

Second, we agree with the insight offered by hard-line compatibilists that what is important for an agent's freedom and moral responsibility is whether the action in question results from the agent's CAS. Yet our condition adds the requirement that the agent's action have its *causal source* in the agent's CAS. As a result, we disagree with the hard-liner about whether manipulated Manny is as free and responsible as merely determined Danny. Whereas the hard-liner maintains that Manny is just as free and responsible as Danny, since his decision issues from his CAS, we claim that there is a principled distinction between them, because Manny's decision has a causal source that extends beyond his CAS, all the way back to Diana and her intentions in designing Manny's zygote in precisely the way that she does. By contrast, Danny's decision has its causal source in the output of his CAS and not in any variable beyond his control, and thus he has free will and is responsible for stealing.

Here, then, is our soft-line response to the Manipulation Argument:

(1) It is relevant to an agent's free will and moral responsibility for performing an action, *A*, whether *A* has its causal source in the agent (specifically, in a variable representing the output of activity in the agent's CAS prior to the agent's *A*-ing)

Footnote 11 continued
a parsimonious account of Frankfurt (1969) cases. In such cases, the Frankfurt intervener (Black) can ensure what an agent (Jones) decides, but when we consider whether Jones is morally responsible in the actual case where Black does nothing, Black is not an actual cause of Jones' decision, and hence Black is not the causal source of that decision. Instead, Jones is.

This premise agrees with the source incompatibilist that causal sourcehood matters for free will and moral responsibility. Yet by contrast with the source incompatibilist, we offer a more precise—and more generally applicable—definition of causal sourcehood, using the resources of the interventionist theory of causation, namely:

(2) $X = x$ is the causal source of $Y = y$ iff X bears the strongest causal invariance relation to Y among all the prior causal variables (including X) that bear such relationships to Y.

Together, (1) and (2) explain *why* it is relevant to Manny's responsibility whether his zygote was intentionally designed by Diana to ensure his stealing, since:

(3) If Manny decides to steal because Diana designs Manny's zygote with the intention of ensuring that he decides to steal, then Diana's decision bears the strongest causal invariance relation to Manny's stealing, among all the prior causal variables that bear such relations to it.

The causal invariance relation obtaining between Diana's prior intention and Manny's stealing the wallet is stronger than the invariance relation obtaining between any other variable and his action of stealing, including the relation between the output of Manny's own deliberations and his stealing. Thus, Diana's deciding in a particular way is the causal source of Manny's action.[12] Yet Danny's situation is different:

(4) If Danny decides to steal, then the output of Danny's prior deliberation bears the strongest causal invariance relation to his stealing, among all the prior causal variables that bear such relations to it.

Why? Because there are no variables prior to Danny's deliberation that bear stronger causal invariance relations with his action. For instance, no other prior variable would cause him to steal across a relevant range of variations in background conditions (see Sect. 5). Thus, the output of Danny's own deliberation is the causal source of his stealing. Given premises (2)–(4), it follows that:

(5) There is a principled relevant difference in whether events in Manny and Danny are the causal sources of their respective actions of stealing—no events in Manny are the causal source of his action, whereas events in Danny are.[13]

[12] The same is true in Pereboom's cases 1 and 2, where neuroscientists can ensure, by manipulating Plum's brain, that his action is caused by their relevant intentions. In case 3, it is less clear how the community that raises Plum could *ensure* his action. To respond to soft-line responses that reference the manipulator's intentions, Pereboom introduces alterations to his cases involving a spontaneously generated machine or force field that produces the same outcome in Plum. Pereboom suggests these alterations are irrelevant to his argument (e.g., 2014: 79), but in these cases the causal invariance relation collapses, since those causes will not produce the action across different background conditions (our condition 2), nor would interventions on the values of the relevant variables (e.g., on the force field) cause systematic changes to Plum's actions (our condition 1) (See Sect. 5 below.).

[13] We noted earlier (footnote 8) that since strength of invariance allows for relative comparisons among causal variables, we could also define *relative* causal sourcehood among selected causal variables. Doing so would alter some of the preceding presentation, but it would still support the conclusion that Diana's

From (1) and (5), we can further conclude that:

(6) There is a principled relevant difference in whether Manny and Danny act freely and are morally responsible for their actions.

Generalizing (as the Manipulation Argument asks us to do from the specific cases it offers), it follows that:

(7) There is a principled difference relevant to free will and moral responsibility for actions between two otherwise identical agents in deterministic universes, one of whom is intentionally manipulated or designed to perform an action and the other of whom is not.

This conclusion demonstrates that premise (2) of the Manipulation Argument—the *NoDif Premise*—is false, and so it undermines the incompatibilist conclusion. There *is* a difference between manipulation and determinism that is relevant to free will and moral responsibility. As a result, compatibilists can reject the most powerful argument currently available to incompatibilists. And they can do so without having to defend the counterintuitive claim that an agent who is intentionally designed by another agent is free and responsible, as hard-line compatibilists insist. Further, they can do so by *drawing on*, rather than *shying away from*, the source incompatibilist's insight that being the causal source of one's actions is an important component of acting freely and responsibly. In Aristotle's terms, there is an external causal source for Manny's action—Diana is "the source and begetter of his action"—whereas for Danny, we "cannot trace back his actions to starting points other than those within him."

5 Replies to objections

When steering a course between a number of opponents, one can expect to be attacked from many angles. Source incompatibilists will say that our causal sourcehood condition is too weak. Hard-line compatibilists will say that it is too strong. Here, we provide further reasons to think that it is just right. In fact, our condition can even tell us when the hard-line response to Manipulation Arguments is the right one to take. However, we first consider a more basic objection targeting the interventionist framework itself. In particular, someone might object that our reliance on interventionism is ill-advised, especially if the view is taken to suggest that the notion of causation that we employ is subjective, whereas the relevant notion should be objective. The worry here is that because the interventionist notion of causation depends on our causal judgments and practices, it is anthropocentric and therefore subjective.

First, the interventionist notion of causation is not subjective, even if it is anthropocentric (see, e.g., Weslake 2006). The causal invariance relations in which

Footnote 13 continued
decision is a relatively stronger causal source of Manny's stealing than any variable in Manny's CAS, whereas there is no causal source in Danny's distant past that is a relatively stronger causal source of his decision than variables in his CAS.

causal models consist express causally sustainable generalizations—in other words, they represent that certain events' occurring would reliably result in certain other events' occurring. Since the models reflect objective causal structure in this way, they are not subjective.

Second, objections claiming that interventionism gives us, at best, an account of our causal practices or judgments, rather than providing (as it should) a theory of what metaphysically causes what, come close to begging an important methodological question. As we saw earlier, interventionism is being used to address a number of issues in metaphysics (see Sect. 3), including issues about causation (e.g., Woodward 2003; Hitchcock 2001). Surely it is an open methodological question (one that we cannot answer here) whether the interventionist approach to causation—as well as to other topics in metaphysics—is the correct one to adopt (e.g., Woodward forthcoming; Kuorikoski 2014). We do not assume that it is, yet we insist that it is a strong contender, which, if correct, supports a soft-line response to the Manipulation Argument.

Leaving aside general worries about the interventionist framework, our argument faces more specific objections. For instance, the source incompatibilist may argue that our account is too weak because, if determinism is true, there will be sufficient causal conditions for the decisions or actions of an agent like Danny, and the agent will lack control over those conditions. Source incompatibilists will define sourcehood as requiring that the agent have control over these conditions (Pereboom 2014: 74). Yet there are problems with defining sourcehood in this way. Most importantly, such a definition appears to be ad hoc, and designed precisely to apply to questions about free will and moral responsibility. If instead it were meant to provide information about causal relations in general—for instance, about which among the causal variables influencing an outcome are more significant than others—it would be useless, since it would entail that all causal relations are equally (in)significant (except perhaps one, the very first cause, if there is one). If causal sourcehood is defined *so that* it contrasts with causal determinism, it should not be surprising that determinism rules out causal sourcehood, thereby supporting an incompatibilist conclusion. Yet this move would seem to beg the question against any possible compatibilist definition of causal sourcehood proposed as relevant to free will and moral responsibility.

By contrast, our causal sourcehood condition is not ad hoc, is useful and fully generalizable, and does not beg any questions. It provides a way of discerning which causal variables are more causally significant than others. Say a computer chip is broken in your car's engine, with the result that an oil clog does not get cleared. Your mechanic will causally model the car's engine in order to locate what she needs to fix so that the clog gets cleared. She must find the type of event occurring within the engine—the relevant causal variable—on which an appropriate intervention would bring about the desired change. Her doing so requires that she identify a robust causal invariance relation between that type of event and the occurrence of the clog. More than that, she must find the *strongest* causal invariance relation obtaining between the variable representing the clog's occurrence and any of its prior causal variables, in order to identify the *causal source* of the problem. Her doing so will enable her to fix the problem. Since the problem lies in the

computer chip, that is where the causal source of the clog lies, and it is where the mechanic should intervene. Alternatively, of course, the mechanic might clear the clog manually, or rig something that bypasses the chip altogether, but still clears the clog. Yet surely we would prefer that the mechanic fix the chip, so that the problem does not simply reoccur, thus requiring her to clear it again or rig another temporary fix.

Our definition of causal sourcehood is neither ad hoc nor designed to apply only to questions about free will and moral responsibility (or manipulation or determinism). It is a fully general principle, and the sort of principle that ordinary people and scientists use all the time in identifying causal explanations.[14]

From the other direction, hard-line compatibilists might think that our causal sourcehood condition is too stringent and thus they might reject our premise (1), the claim that it is relevant to an agent's free will and moral responsibility for performing an action whether that action has its source in the agent's CAS. For these compatibilists, our condition may suggest that people are free and responsible for too few actions, since many actions might appear to have stronger causal invariance relations with events occurring *outside* an agent's CAS than they do with appropriate events occurring *within* the CAS. There are two possible worries here. First, one might worry that some *diachronic* causal variable in Danny's distant past (and thus external to his CAS) counts as the causal source of his decision. Second, one might worry that some external *synchronic* causal variable at, or just before, the time of his decision is the causal source.

Here is our response to the first worry. When scientists seek to causally explain why a given event outcome occurs, they are rarely interested in what would have happened if an event 30 years prior to the outcome event had occurred differently than it did (except in cases where they are modeling relatively stable systems in which they can keep fixed relevant background conditions; for instance, the orbits of Halley's comet in the gravitational conditions of our solar system). In part, this is because typically there is no way to model how systematic changes in the values of the variable representing the event occurring 30 years ago might systematically relate to changes in the later event. According to interventionism, a causal variable functions as an epistemic tool that enables us (ideally, at least) to make changes to the type of event we are explaining. We do not, of course, actually need to be able to perform the required changes; it is enough that we can make sense of how outcomes would change, were we to make certain interventions. In this procedure, appealing to variables in the distant past is usually unhelpful, since we have no way of

[14] Recent studies indicate that ordinary causal judgments track the distinctions that our causal sourcehood condition suggests. In studies about manipulation cases, when people read that a manipulator causes another agent to do what she intends, they judge her to be the cause of, and responsible for, the relevant action (or to be more so than the manipulated agent), whereas people do not make those judgments (or not as strongly) when the manipulator causes another agent to do the same thing, but without the manipulator intending it (e.g., Murray and Lombrozo 2016; Phillips and Shaw 2014; Sripada 2012).

modeling how an intervention on such an event would systematically change another event in the present that we are trying to explain.[15]

While determinism says that some immense set of events occurring in the distant past, together with the laws, cause events occurring now with probability 1, we have no way of knowing how altering some event to occur in one way rather than another would causally explain why a particular event rather than another event happens much later. Of course, if the Earth had been destroyed 30 years ago, that would affect Danny's decision to steal, since it would prevent Danny from existing. Yet such a hypothetical intervention cannot help to explain why Danny decides to steal rather than return the wallet. It is implicit in the interventionist framework that causal invariance relations describing an intervention specify contrastively causal relations.[16] Yet causal relations of this sort cannot be specified in Danny's case, if the antecedent event occurred in the distant past. By contrast, a hypothetical intervention on the variable representing Diana's decision *does* result in a change to the value of the variable representing Manny's stealing, under the specified conditions, which in turn explains why Manny decides to steal rather than return the wallet. Further, it would not help to expand the set of events on which we might intervene in Danny's case to include the entire state of the universe at some time in the distant past, given that alterations to that entire state (or parts of it) would not alter later events in any contrastively specifiable way either.[17]

There remains the worry that some synchronic causal variable at, or just before, the time of Danny's decision, which lies outside his CAS, is the causal source of his decision. Here, our causal sourcehood condition reveals a potential limitation to free and responsible agency. If a stronger causal invariance relation obtains between, say, the occurrence of ambient lawnmower noise and an agent's walking past someone in need without helping them (Matthews and Canon 1975) than obtains between events occurring within the agent's CAS and her subsequent action, then the causal source of the action is the noise. In such cases, our causal sourcehood condition might reveal that agents sometimes have at least reduced responsibility for their actions, because the causal source of their action lies (at least partly; see footnotes 6 and 8) outside their CAS. We take this consequence to be an advantage of our view, since it permits a way of mapping degrees of freedom and moral responsibility in light of empirical discoveries about the relative strengths of different causal factors on our decisions and actions.

[15] An exception, of course, is for causal variables like Diana's decision, since Diana is stipulated to be capable of controlling for a maximally wide range of possible changes to the background conditions (something we doubt is nomologically possible).

[16] In fact, there may be need for *explicit* contrastivity (Deery 2013); however, we do not have space to develop that point here.

[17] Even if a suitable invariance relation *could* be identified between some event(s), X, in the distant past and Danny's decision (*Danny*), our conditions on strength of causal invariance would show that there would rarely, if ever, be a *strong* causal invariance relation between X and *Danny*, since $X = x$ would not cause *Danny* = d across a wide range of changes to C.

Indeed, our causal sourcehood condition also provides a useful metric indicating when compatibilists should *switch* from the soft-line to the hard-line response, and vice versa. Compatibilists need principled reasons to determine when to reject the *NoDif Premise* of the Manipulation Argument and when instead to reject the *NoFW Premise*. Soft-line responses have typically argued that manipulation differs from determinism by stipulating that manipulation is intuitively freedom-undermining in a way that determinism is not (a response that risks begging the question) or that a manipulated agent like Manny is less morally responsible because another agent (Diana) is responsible for his actions. Advocates of Manipulation Arguments have responded by presenting cases in which there is *no* manipulating agent, or the manipulator cannot be held morally responsible (e.g., because she is insane; see Mele 2013). Our argument does not rely on the intuition that responsibility or blame gets "transferred" to the manipulator (although it allows that Diana might be *more* responsible than Manny). Rather, our soft-line response is based on the manipulator's being an intentional agent with control over her intended outcome, such that her decision is the causal source of that outcome.

However, if an advocate of the argument "waters down" the cases so that there is no intentional manipulator, but just a factor that is stipulated to causally influence an agent's later decision, then our metric of causal sourcehood explains why the hard-line response is appropriate in those cases. Diana is the causal source of Manny's action if, and only if, she intends to cause that action and has the knowledge and power to ensure that it occurs, because only then is there a stronger invariance relation between her decision and Manny's action than there is between the output of Manny's deliberative activity and the action. If instead Diana is described as merely getting lucky in creating Manny such that he steals the wallet, or if Diana is replaced with "force fields or machines that randomly form in space" (Pereboom 2014: 82) that somehow affect the creation of Manny's zygote, but with no control over what he later does, then there is a very weak invariance relation between these "manipulators" and Manny's action. In these cases, Diana or her non-agential replacement is *not* the causal source of Manny's action. Unless one were *already* worried that the deterministic features of these cases challenge free will or responsibility (in which case the argument itself does no work), one should recognize that these cases of non-agential causal factors do not challenge Manny's being the causal source of his stealing. Indeed, evidence suggests that most people *do* recognize that the manipulator's intentions make a causal difference to the manipulated agent's actions, one that is relevant to that agent's causal role in, and moral responsibility for, those actions (see footnote 14).

6 Conclusion

Debates about free will, moral responsibility, and determinism depend crucially on facts about causal relations and causal sourcehood. Source incompatibilists and their Manipulation Arguments have helped to focus attention on these issues. We have argued that interventionist causal modeling is useful for analyzing claims about causal relations as they feature in debates about free will. The notion of causal

invariance enables us to develop a fully general definition of causal sourcehood that applies not only to debates about free will and moral responsibility. Using this definition, we can see why agents who are manipulated (or designed) by a powerful agent to act in accordance with the manipulator's intentions are not the causal sources of their actions. Conversely, determinism alone does not entail that agents cannot be the causal sources of their actions. Indeed, we can see why, even if determinism is true, agents' exercising their compatibilist capacities (like rational deliberation and self-reflective identification) can be the causal source of their subsequent actions, since no earlier variables bear a stronger causal invariance relation with these actions than the variables describing the output of the relevant capacities.[18]

In this way, we agree with advocates of the Manipulation Argument that causal sourcehood is important for free will and moral responsibility. Yet our compatibilist sourcehood condition blocks the incompatibilist conclusion, supporting a soft-line compatibilist response to the argument. Furthermore, our sourcehood condition explains why, in other sorts of cases problematically presented as manipulation, the hard-line compatibilist is correct to insist that the "manipulated" agents can indeed be free and morally responsible for their actions.[19]

References

Aristotle. 350 BC/2011. *Nicomachean ethics* (R. C. Bartlett & S. D. Collins, Trans.). Chicago: University of Chicago Press.

Ayer, A. J. (1954). Freedom and necessity. *Philosophical essays* (pp. 271–284). London: Macmillan.

Barnes, E. C. (2013). Freedom, creativity, and manipulation. *Noûs, 49*(3), 560–588.

Campbell, J. (2010). Control variables and mental causation. *Proceedings of the Aristotelian Society, 110*, 15–30.

Capes, J. (2013). Mitigating soft compatibilism. *Philosophy and Phenomenological Research, 87*(3), 640–663.

Cartwright, N. (1979). Causal laws and effective strategies. *Noûs, 13*, 419–437.

Coates, J., & Swenson, P. (2013). Reasons-responsiveness and degrees of responsibility. *Philosophical Studies, 165*, 629–645.

Deery, O. (2013). Absences and late preemption. *Theoria, 79*(4), 309–325.

[18] We also maintain that interventionism can be fruitfully applied to other debates about free will. For instance, it can explain why the agent in a Frankfurt case (1969) is the causal source of her action whereas the counterfactual intervener is not, offering a more parsimonious explanation of the lessons of those cases (see footnote 11). Additionally, interventionism can help to explain away features of our experience of free agency that appear to implicate indeterminism (Deery 2015). We hope to further develop these applications of interventionism in future work.

[19] Versions of this paper were presented at Duke University (March 22, 2013), the University of Fribourg, Switzerland (June 18, 2013), the *Flickers of Freedom* blog (July 2013), the University of Montreal (November 29, 2013), the 40th Annual Meeting of the Society for Philosophy and Psychology (June 20, 2014), and the University of Nevada, Las Vegas (October 24, 2014). Thanks to audiences at those venues for helpful comments. We also thank Terry Horgan, Derk Pereboom, Al Mele, Michael McKenna, Jenann Ismael, Carolina Sartorio, Adina Roskies, Paul Russell, Jonathan Phillips, Dylan Murray, Bryan Chambliss, Alex Von Stein, Yael Lowenstein, and several anonymous referees for their helpful comments and suggestions.

Deery, O. (2015). Why people believe in indeterminist free will. *Philosophical Studies, 172*(8), 2033–2054.
Demetriou, K. (2010). The soft-line solution to Pereboom's four-case argument. *Australasian Journal of Philosophy, 88*(4), 595–617.
Dennett, D. (1984). *Elbow room: The varieties of free will worth wanting*. Cambridge, MA: MIT Press.
Fischer, J. M. (2000). Responsibility, history and manipulation. *The Journal of Ethics, 4*(4), 385–391.
Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge: Cambridge University Press.
Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy, 66*(23), 829–839.
Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy, 68*(1), 5–20.
Haji, I., & Cuypers, S. (2006). Hard- and soft-line responses to Pereboom's four-case manipulation argument. *Acta Analytica, 21*(4), 19–35.
Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy, 98*(6), 273–299.
Ismael, J. (2013). Causation, free will, and naturalism. In H. Kincaid, J. Ladyman, & D. Ross (Eds.), *Scientific metaphysics* (pp. 208–235). New York: Oxford University Press.
James, W. (1890). *The principles of psychology*. Cambridge, MA: Harvard University Press.
Kane, R. (1996). *The significance of free will*. New York: Oxford University Press.
Khoury, A. (2014). Manipulation and mitigation. *Philosophical Studies, 168*(1), 283–294.
Kuorikoski, J. (2014). How to be a humean interventionist. *Philosophy and Phenomenological Research, 89*(2), 333–351.
Lagnado, D., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science, 37*, 1036–1073.
List, C., & Menzies, P. (2009). Nonreductive physicalism and the limits of the exclusion principle. *Journal of Philosophy, 106*, 475–502.
Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology, 61*(4), 303–332.
Matthews, K. E., & Canon, L. K. (1975). Environmental noise level as a determinant of helping behavior. *Journal of Personality and Social Psychology, 32*(4), 571–577.
McCain, K. (2012). The interventionist account of causation and the basing relation. *Philosophical Studies, 159*, 357–382.
McKenna, M. (2008). A hard-line reply to Pereboom's four-case manipulation argument. *Philosophy and Phenomenological Research, 77*(1), 142–159.
McKenna, M. (2014a). Compatibilist ultimacy: Resisting the threat of Kane's U-condition. In D. Palmer (Ed.), *Libertarian free will: Contemporary debates* (pp. 71–87). Oxford: Oxford University Press.
McKenna, M. (2014b). Resisting the manipulation argument: A hard-liner takes it on the chin. *Philosophy and Phenomenological Research, 89*(2), 464–484.
Mele, A. (1995). *Autonomous agents: From self-control to autonomy*. New York: Oxford University Press.
Mele, A. (2006). *Free will and luck*. New York: Oxford University Press.
Mele, A. (2013). Manipulation, moral responsibility, and bullet biting. *Journal of Ethics, 17*(3), 167–184.
Menzies, P. (2004). Causal models, token causation, and processes. *Philosophy of Science, 71*(5), 820–832.
Mickelson, K. (2015). The zygote argument is invalid: Now what? *Philosophical Studies, 172*(11), 2911–2929.
Murray, D., & Lombrozo, T. (2016). Effects of manipulation on attributions of causation, free will, and moral responsibility. *Cognitive Science*. doi:10.1111/cogs.12338.
Nelkin, D. (2016). Difficulty and degrees of moral praiseworthiness and blameworthiness. *Noûs, 50*(2), 356–378.
O'Connor, T. (1995). Agent causation. In T. O'Connor (Ed.), *Agents, causes, and events: Essays on indeterminism and free will* (pp. 173–200). New York: Oxford University Press.
Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.
Pereboom, D. (2001). *Living without free will*. Cambridge: Cambridge University Press.
Pereboom, D. (2008). A hard-line reply to the multiple-case manipulation argument. *Philosophy and Phenomenological Research, 77*(1), 160–170.
Pereboom, D. (2014). *Free will, agency, and meaning in life*. Oxford: Oxford University Press.

Phillips, J., & Shaw, A. (2014). Manipulating morality: Third-party intentions alter moral judgments by changing causal reasoning. *Cognitive Science, 39*(6), 1320–1347.

Ragland, C. P. (2011). Softening Fischer's hard compatibilism. *The Modern Schoolman, 88*(1/2), 51–71.

Roskies, A. (2012). Don't panic: Self-authorship without obscure metaphysics. *Philosophical Perspectives, 26*(1), 323–342.

Sartorio, C. (2013). Making a difference in a deterministic world. *Philosophical Review, 122*(2), 189–214.

Schlosser, M. (2015). Manipulation and the zygote argument: Another reply. *Journal of Ethics, 19*(1), 73–84.

Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives*. New York: Oxford University Press.

Sripada, C. (2012). What makes a manipulated agent unfree? *Philosophy and Phenomenological Research, 85*(3), 563–593.

Tierney, H. (2013). A maneuver around the modified manipulation argument. *Philosophical Studies, 165*(3), 753–763.

Todd, P. (2011). A new approach to manipulation arguments. *Philosophical Studies, 152*(1), 127–133.

Van Inwagen, P. (1983). *An essay on free will*. Oxford: Oxford University Press.

Vihvelin, K. (2013). *Causes, laws, and free will: Why determinism doesn't matter*. New York: Oxford University Press.

Waller, R. R. (2014). The threat of effective intentions to moral responsibility in the zygote argument. *Philosophia, 42*, 209–222.

Watson, G. (2000). Soft libertarianism and hard compatibilism. In M. Betzler & B. Guckes (Eds.), *Autonomes Handeln: Beitrage zur Philosophie von Harry G. Frankfurt* (pp. 59–70). Berlin: Akademie Verlag.

Weslake, B. (2006). Review of James Woodward, *Making things happen*. *Australasian Journal of Philosophy, 84*(1), 136–140.

Wolf, S. (1987). Sanity and the metaphysics of responsibility. In F. Schoeman (Ed.), *Responsibility, character and emotions: New essays on moral psychology* (pp. 46–62). Cambridge: Cambridge University Press.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. New York: Oxford University Press.

Woodward, J. (2015). Interventionism and causal exclusion. *Philosophy and Phenomenological Research, 91*(2), 303–347.

Woodward, J. (Forthcoming). Interventionism and the missing metaphysics: A dialogue. In M. Slater and Z. Yudell (Eds.), *Metaphysics and the philosophy of science*. Oxford: Oxford University Press.

Yang, E. (2013). Eliminativism, interventionism and the overdetermination argument. *Philosophical Studies, 164*, 321–340.