Desperately Seeking Sourcehood

Hannah Tierney, University of Sydney, hannah.tierney@sydney.edu.au

David Glick, University of Sydney, david.glick@sydney.edu.au

Abstract

In a recent essay, Oisín Deery and Eddy Nahmias (2017) utilize interventionism about causation to develop an account of causal sourcehood in order to defend compatibilism about free will and moral responsibility from manipulation arguments. In this paper, we criticize Deery and Nahmias' analysis of sourcehood by drawing a distinction between two forms of causal invariance that can come into conflict on their account. We conclude that any attempt to resolve this conflict will either result in counterintuitive attributions of moral responsibility or will undermine their response to manipulation arguments.

1. Introduction

Interventionism, roughly, is the view that one variable is the cause of another variable iff an intervention on the former would lead to a change in the latter.[1] Interventionism has had a transformative effect on discussions of causation in computer science, psychology, and several areas of philosophy. So, it is perhaps unsurprising that philosophers have begun to explore the ways in which interventionism can shed light on issues surrounding the free will debate. For example, Jenann Ismael (2013, 2016) utilizes an interventionist approach to causation to defuse the threat of determinism to free will. And Joseph Campbell (2010) develops an interventionist account of control variables in order to provide an explanation of mental causation, while Adina Roskies (2012) uses Campbell's framework to construct a view of self-authorship. Additionally, philosophers have recently relied on interventionism to defend compatibilism from manipulation arguments. Oisín Deery and Eddy Nahmias (henceforth DN) (2017) formulate an interventionist theory of causal sourcehood,

[1] This view has been developed and defended primarily by Judea Pearl (2009) in computer science and James Woodward (2003, 2016) in philosophy.

while Marius Usher (2018) presents an interventionist account of teleological control, in response to manipulation arguments.

In this essay, we focus on the recent attempts to overcome manipulation arguments that rely on interventionism, and specifically on DN's account of causal sourcehood.[2] There are several reasons for this. First, DN argue that their account of causal sourcehood can rescue compatibilists from the particularly difficult position manipulation arguments have left them in. As John Martin Fischer notes, "…manipulation cases are compatibilism's dirty little secret…We compatibilists have to deal with this" (Fischer 2000: 390). DN also argue that their response to manipulation arguments goes further and is more systematic than other accounts (2017: 1260) and should thus attract the attention of both compatibilists and their opponents alike. Second, though DN propose their account of causal sourcehood in a very specific context, it is meant to be fully general. This gives their account a certain broadness of scope—it applies not only to cases featuring manipulation and questions about morally responsibility, but also to completely unrelated areas where we may be interested in determining the causal sources of events. In this way, DN's account of causal sourcehood has both substantial strength and breadth—it bears on a number of philosophical debates and will be of interest to a broad range of philosophers.

Our plan for the paper is as follows. In section 2 we examine DN's account of causal sourcehood and their response to manipulation arguments. Then, in section 3, we draw a distinction between two forms of invariance on their account that can come into conflict. In section 4 we argue that any attempt to resolve this conflict will either result in counterintuitive attributions of moral responsibility or will undermine DN's response to manipulation arguments. Next, in section 5, we

[2] Though we focus on DN's view of causal sourcehood in this essay, many of our criticisms apply to Usher's (2018) account of teleological control as well, which we highlight in footnote 15. We also criticize one of Usher's objections to DN in footnote 5.

consider a possible rejoinder on behalf of DN and argue that this reply renders their account either circular or uninformative. Finally, in section 6, we conclude that while DN's objection to manipulation arguments is ultimately unsuccessful, interventionism is not without import for matters of moral responsibility.

2. DN's Soft-Line Response to Manipulation Arguments

DN (2017) focus their discussion of manipulation arguments on a set of cases developed by Alfred Mele (2013):

> 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. 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. (DN 2017: 1257)

Compare this to a slightly different case:

> …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 thirty 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 thirty 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. 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). (DN 2017: 1257)

Proponents of manipulation arguments contend that Manny intuitively does not have free will and is not morally responsible for stealing the money. More generally, they argue that any agent who is manipulated in the way that Manny is does not have free will. DN call this more general claim NoFW.

Proponents of manipulation arguments also argue that Manny and Danny meet all the same compatibilist sufficient conditions for free will.[3] And, they claim that, intuitively, there is no difference between the kind of manipulation featured in the case of Manny and the truth of determinism, at least when it comes to free will and moral responsibility. DN call this thesis NoDif. If both NoFW and NoDif are true, then Danny, and all agents who live in deterministic universes, are neither free nor morally responsible, just like Manny. Thus, the defender of the manipulation argument can conclude that incompatibilism is true.

There are many avenues to pursue when responding to a manipulation argument. One could, for example, deny NoFW and argue that Manny and those manipulated like him are in fact both free and morally responsible. One could also deny NoDif and argue that there is a relevant difference between manipulation and determinism such that even though manipulated agents like Manny aren't free and morally responsible, determined agents can be. Michael McKenna (2008) calls the strategy of denying NoFW taking the "hard-line" while those who deny NoDif take the "soft-line." McKenna (2008) favors taking the hard-line, since soft-line responses tend to be mere stop-gap solutions that only keep the incompatibilist at bay temporarily. According to McKenna (2008), for any novel condition the compatibilist proposes that the determined agent meets but the manipulated agent does not, it will be in principle possible for the defender of the manipulation argument to augment the cases such that the manipulated agent meets the proposed condition in question as well. But while taking the soft-line may only succeed temporarily, taking the hard-line is no easy path; it requires the compatibilist to defend the, perhaps counterintuitive, claim that Manny and those manipulated like him do in fact have free will and are morally responsible for the actions for which they were

[3] DN (2017) point to Fischer and Ravizza (1998), Frankfurt (1971), Wolf (1990), and Mele (1995) for different accounts of the minimally sufficient compatibilist conditions for free will.

manipulated to perform. Thus, the compatibilist is in a difficult position—the most principled and stable response to manipulation arguments also tends to be the most counterintuitive.

Happily, DN argue that it's possible for soft-line approaches to manipulation arguments to be more than temporary roadblocks for incompatibilists to overcome. DN, relying on interventionist approaches to causation, develop an account of causal sourcehood that allows them to reject NoDif and argue that there is an in-principle difference between Manny and Danny (and more generally, between manipulated and determined agents). They also contend that there is no (easy) way for the defender of manipulation arguments to alter their cases to shore up NoDif in response to their interventionist objection.

Interventionism about causation involves constructing models that represent counterfactual relationships among events. If we want to know whether the output of Danny's (or Manny's) Compatibilist Agential Structure (CAS) caused him to steal the money, we first assign variables to the outputs of his CAS (X) and the stealing of the money (Y).[4] Next, we can consider an intervention to determine whether X caused Y. This involves setting X to a different value and seeing whether it changes the value of Y. For example, we can imagine that the output of Danny's CAS was the decision not to steal the money (X=0). In this case, the intervention would likely result in a change in the value of Y, i.e. the stealing of the money wouldn't occur (Y=0).

On an interventionist approach to causation, the output of both Danny's and Manny's CASs count as actual causes of their stealing the money. This is because changes in the value of the variables representing the output of their CASs result in a change to the value of the variables representing their stealing the money. Though the outputs of Danny's and Manny's CASs are both actual causes of their respective acts of stealing the money, DN argue that there is a relevant different between how Danny

[4] DN argue that CAS represents "…the features of an agent's psychology that compatibilists typically judge as jointly (and minimally) sufficient for free will and moral responsibility" (DN 2017: 1258).

and Manny causally relate to their actions. While Danny's CAS is the *causal source* of the event, Manny's is not. And, DN contend, an agent cannot be free or morally responsible with respect to a given action if the agent's CAS was not the causal source of that action (2017: 1267).

According to DN, for a variable to be the causal source of an event, that variable must bear the strongest causal invariance relation to it among all the variables that are causally connected to the event:[5]

> 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. (DN 2017: 1262–1263)

In the case of Danny, the output of his CAS is plausibly the causal source of his stealing the money. Changes in the value of the variable representing the output of Danny's CAS result in a change to the value of the variable representing Danny's stealing the money. And because Danny is an intentional agent, his decision to steal the money would lead to him doing so across many changes in the background conditions. For example, Danny would likely steal the money if it were raining or snowing, if the wallet was black or brown, etc. But there will be some changes in the background conditions that would result in Danny not stealing the money, e.g., the sudden appearance of a police officer.

[5] DN allow for more than one variable to bear the strongest invariance relation to an event and hence qualify as the causal source of that event. In such a case, each agent whose CAS is a causal source of an event would be in a position to be morally responsible for that event. They argue: "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" (DN 2017: 1263, footnote 8).

Much of this is true of Manny as well. The relationships between the variables representing the outputs of both Danny's and Manny's CAS and the variables representing their stealing the money are equally invariant. But there is a stronger causal invariance relation between Diana's meddling and Manny's stealing, according to DN. Importantly, Diana intends for Manny to steal the money at t_{30} and she is able to ensure that he does so. So, there's no change (or very few changes) in the background circumstances that could break the causal connection between Diana's decision to meddle and Manny stealing the money.[6] But, when evaluating the relationship between the output of Manny's CAS and his stealing the money, there are a number of changes in the background circumstances that would break the link between the variables. When evaluating this relationship, we must ignore Diana and all other causally relevant events—we are only concerned with the output of Manny's CAS and his stealing the money. And just as there are changes to the background circumstances that would stop Danny from stealing the money, e.g., the sudden appearance of a police officer, these changes would stop Manny as well. Thus, Diana's decision relates more invariantly to Manny's stealing the money, and hence she is the causal source of this event, not Manny.

In developing an account of causal sourcehood, DN claim to have isolated a relevant difference between Danny and Manny—while the output of Danny's CAS is the causal source of his

[6] Interestingly, Usher (2018), whose response to manipulation arguments also relies on an interventionist approach to causation, defends a hard-line response to the Diana case. He claims that because Diana acts 30 years before the theft occurs, "…if we are to make any intervention on background circumstances (say bringing about rain on the day of the theft), there is nothing that Diana could do to "ensure" that Manny steals the wallet" (2018: 19). But this is to mistake a change in background conditions with an intervention. To assess whether Diana is an actual cause, we consider an intervention—a surgical change in the value of her decision—and ask whether this would change the value of the variable representing the theft. Once we've established that Diana is an actual cause, we can assess the invariance of the causal relation by asking under what range of background conditions it will continue to obtain. To do this we consider a possible world in which some non-causal features are different—say, it's raining at t_{30}—and ask whether the causal relation still obtains. That is, we consider whether an intervention on Diana's decision in that world successfully predicts whether or not the money is stolen. In the present case, it seems that it would—Diana would simply implement her plan somewhat differently to ensure that Manny isn't dissuaded by the rain. Thus, to us, DN's claim that Diana's decision is the most invariant cause of Manny's stealing the money is plausible. If this is so, then Usher should agree with DN that Diana's presence at least mitigates Manny's responsibility.

stealing the money, the output of Manny's CAS is not. And, DN argue that in order to be free and morally responsible with respect to a particular action, an agent's CAS must be the causal source of that action (2017: 1267). So, while Danny is in the position to be free and morally responsible with respect to his stealing the money, Manny is not. If correct, DN can reject NoDif both as it pertains to Danny and Manny and the more general claim that there is no difference between the kind of manipulation featured in the case of Manny and the truth of determinism when it comes to free will and moral responsibility.[7] While manipulation renders agents unable to be the causal sources of their actions, and thus not in a position to be morally responsible for them, determinism does no such thing.

3. Invariance, Reliability, and Stability

DN have developed an account of causal sourcehood that is both independently motivated and applies well beyond the bounds of the free will and moral responsibility literature. However, their account of causal sourcehood combines two forms of invariance that are conceptually distinct and can come apart in interesting ways.[8] We follow Woodward (2007: 76-77) in calling invariance under changes of background conditions *stability* (DN's condition 2).[9] And we refer to invariance under changes of the

[7] In light of DN's soft-line response, a proponent of manipulation arguments may revise the manipulation case such that Diana cannot *ensure*, or perhaps even intend, that Manny will go on to steal the money. Given these revisions, it's unlikely that Diana's meddling would relate most invariantly to Manny's stealing the money. In response to this modified version of the Diana case, DN argue that they would take the hard-line: if Diana's decision does not relate most invariantly to Manny's stealing the wallet, then Diana cannot be the causal source of this event and Manny, like Danny, is in the position to be morally responsible for the theft (2017: 1273). However, as DN note, many defenders of manipulation arguments are quick to make such modifications to their manipulation cases, replacing the manipulator with a spontaneously generated machine or force field (Pereboom 2001, 2014; Mele 2005). Given this, it's hard to see how DN's account provides a more successful and systematic response to manipulation arguments than other soft-line responses. But, rather than object to DN on these grounds, we will focus on their positive account of causal sourcehood.

[8] Woodward (2003) also runs these together, but in later work distinguishes them.

[9] See also, Woodward (2006), which contains an extensive discussion of stability in terms of *insensitivity* to background conditions. And Usher (2018) refers to invariance under changes to background conditions as *robustness*.

value of the causal variable as *reliability* (DN's condition 1).[10] So, for DN, a variable X, related by R_1 to Y, is the *causal source* of Y iff for any other causal variable W, related by R_2 to Y, R_2 is no more reliable and stable than R_1, where reliability and stability are determined in the following way:[11]

> **[Reliability]:** 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
> **[Stability]:** R_1 predicts the value of Y across a wider range of relevant changes to the values of C than R_2 does. (DN 2017: 1262–1263)

Once these different forms of invariance are distinguished, one might wonder how they should be weighed against one another in cases where they indicate different causal sources. For instance, suppose variables A and B are both actual causes of C, but A → C is more reliable than B → C while B → C is more stable than A → C. Is A or B the causal source of C? This question becomes even more vexed if A and B are agents' CASs and we want to know who is morally responsible for C.

Imagine that a mafia boss decides to hire an assassin to kill an enemy. The assassin takes the job and decides to kill the enemy by poisoning the enemy's dinner on a Thursday evening and then goes on to do just this. Using DN's analysis, we can model the relationship between the boss's decision (variable B, value b), the assassin's decision (variable A, value a), and the assassin's murder of the enemy (variable M, value m).

$$B{=}b \rightarrow A{=}a \rightarrow M{=}m$$

$$A{=}a \rightarrow M{=}m$$

What is the causal source of the murder: the boss's decision or the assassin's decision?[12] On the one hand, the relationship between the assassin's decision and the murder (R_1) appears to be more reliable

[10] A causal relation which meets DN's condition 1 supports many counterfactuals of a certain type (those that hold fixed the background conditions), hence, it is *reliable.* Here we part company with Woodward (2007), who refers to this simply as *invariance*. For us, invariance is the genus of which stability and reliability are species.

[11] We've rephrased DN's account both to distinguish between reliability and stability and to make clear that multiple variables can each be causal sources of an event. That is, causal sourcehood admits of ties (see footnote 5).

[12] For ease of exposition, we will write in terms of agents' decisions as opposed to the outputs of agents' CASs.

than the relationship between the boss's decision and the murder (R_2). This is because there is a narrower range of values that the boss's decision can take that will predict the murder of the enemy as compared to the range of values that the assassin's decision can take. While the boss can plausibly decide only whether to hire an assassin and which assassin to hire, we can assign a much broader range of values to the assassin's decision.[13] For example, we can imagine a number of different ways the assassin could decide to kill the enemy: by different means, at a different time, on a different day, etc. Thus, holding fixed the relevant background conditions, C, R_1 predicts the value of M under a wider range of interventions on A than R_2 does under interventions on B.[14]

On the other hand, R_2 is more stable than R_1. By hiring the assassin to kill the enemy, the boss provides the assassin with an incentive to perform the murder. The assassin would most likely successfully commit the murder under a wide range of changes in the background circumstances because she was hired to do so. For example, we can easily imagine the assassin successfully following through on the mafia boss's orders if it suddenly began raining outside, or if she came down with a cold, or if it was announced that her favorite movie was playing on TV. But when evaluating the relationship between only A and M, we must ignore the contribution of B. Without the contribution of the boss's orders, there are far more circumstances in which the assassin's decision could be thwarted. If the assassin decided to murder the individual for fun or for practice, then a sudden change in the weather, her health, or access to less violent entertainment could all cause the assassin to not act on her decision. But these are precisely the kinds of changes in circumstances in which the boss's

[13] One might argue that the boss's decision could take a wider range of values than we claim here and still predict the value of the murder. We address this possibility in section 4.c.

[14] It is, of course, difficult to determine the appropriate range of values for variables such as these. The range of values the variables can take will be sensitive to the level of description one adopts, which depends on a variety of contextual factors. While this fact makes it difficult to develop a case in which one agent's decision clearly and uncontroversially relates more reliably to an event than another agent's decision, it also complicates matters for accounts of causal sourcehood that invoke reliability, especially those like DN's that are meant to ground free will and moral responsibility.

orders would still carry weight. Furthermore, if the assassin attempted to murder the enemy, but failed due to a change in the background conditions, the boss could go on to hire as many assassins as is necessary to complete the murder. Thus, there are plausibly fewer circumstances in which the assassin's decision would lead to the murder of the enemy than circumstances in which the boss's decision would. So, R_2 predicts the value of M across a wider range of relevant changes to the background conditions than R_1.

But if R_1 is more reliable than R_2 and R_2 is more stable than R_1, which relation is more invariant on DN's view? DN give us no way to weigh reliability and stability against one another when determining the invariance of causal relationships, and without such a mechanism it is impossible to determine which agent's decision, the boss's or the assassin's, is the causal source of the murder.

4. Stability vs. Reliability

In this section, we explore several potential ways of resolving the conflict between reliability and stability in DN's account of causal sourcehood. We argue that each attempt at resolution either produces counterintuitive judgments about free will and moral responsibility or undermines DN's soft-line response to manipulation arguments.

4.a. No causal source

DN could argue that because neither the boss's decision to have his enemy murdered nor the assassin's decision to murder the boss's enemy relate the most reliably and most stably to the murder of the enemy, neither the boss's decision nor the assassin's decision is the causal source of the murder. Their account of causal sourcehood indicates that a variable is the causal source of an event iff that variable bears the most reliable *and* most stable relationship to that event. But if neither the boss's decision nor the assassin's decision is the causal source of the murder, then neither the boss nor the assassin could be morally responsible for the murder on DN's view.

This strikes us as counterintuitive. Surely at least one of these individuals is morally responsible for the murder. This is not a manipulation case nor is it a case in which determinism need be true. Any theorist who thinks moral responsibility is possible ought to be able to accommodate the intuition that at least one of these individuals is morally responsible for the murder. Agents often conspire with one another to commit crimes. And sometimes, these agents perform different tasks or occupy different roles such that one agent relates more stably to the crime while the other agent relates more reliably. It's difficult to see why this kind of conspiracy renders both agents unfree and absolves them of blame. Furthermore, it's odd to think that if either the boss or assassin had decided to kill the enemy on their own, then their decision would have been the causal source of the murder and they would be open to being morally responsible for it. For these reasons, we think DN's account of causal sourcehood should be modified—either causal sourcehood can be attained without both the reliability and stability conditions being met or causal sourcehood shouldn't be a necessary condition for free will and moral responsibility. We explore both of these possibilities below.

4.b. Two causal sources

Perhaps the most natural response to the boss and assassin case is to think that both agents are in the position to be morally responsible for the murder. To capture this intuition, one could argue that both the boss's decision and the assassin's decision are causal sources of the enemy's murder. It's possible to modify DN's account of causal sourcehood to produce this result by making it disjunctive: A variable X, related by R_1 to Y, is the *causal source* of Y iff for any causal variable W, related by R_2 to Y, R_2 is *either* no more reliable *or* no more stable than R_1. On this account, both the assassin's decision and the boss's decision qualify as causal sources of the murder because the assassin's decision most reliably relates to the murder while the boss's decision most stably relates to it.

Notice, however, that DN can't accept this disjunctive notion of causal sourcehood without conceding that both Manny and Diana are causal sources of Manny stealing the money. Though

Diana's decision relates more stably to the theft, Manny's decision relates just as reliably to the theft as Diana's. DN are aware that Manny and Diana relate equally reliably to the stealing of the money and argue that: "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 [Diana] or instead *Manny* is more strongly invariant regarding Steal" (DN 2017: 1264, footnote 10). So, Manny, by fulfilling the reliability condition, would also count as a causal source of the theft on the disjunctive account.

But if Manny is a causal source of his stealing, then DN's soft-line response to manipulation arguments is jeopardized. DN reject NoDif on the grounds that while Danny is the causal source of his stealing, Manny is not. But on a disjunctive account of causal sourcehood, both Manny and Danny would qualify as causal sources of their stealing—there would no longer be a relevant difference between the two agents. And while DN may be happy to adopt the disjunctive view of causal sourcehood and put forth a hard-line response to manipulation arguments, this will require them to abandon their current approach.

4.c. One causal source

Alternatively, DN could argue that when no variable relates both most reliably and most stably to an event, we should give preference to one of the criteria over the other. If DN were to favor reliability over stability when the two conflict, this would entail that only the assassin is the causal source of the murder and is thus in the position to be morally responsible for the crime. But this strikes us as counterintuitive. It seems plausible that the boss is just as morally responsible for the crime as the assassin. But even if one doesn't share this intuition, it's puzzling that the sheer number (or range) of ways the assassin can bring about the murder is decisive in determining her status as the causal source of, and the morally responsible agent for, the crime. Furthermore, if we imagine that the boss's decision can take a wider range of values than originally supposed (while still predicting the murder), it's also odd that this fact alone may render the boss the causal source of the murder and thus open to blame for the crime. While examining the reliability of an invariance relation between two variables can tell us many things about the causal relationship between them, it's difficult to see how reliability alone can settle questions about causal sourcehood and moral responsibility.[15]

Perhaps DN would be more amenable to favoring stability when the two forms of invariance conflict. After all, they focus almost exclusively on the stability condition in their discussion of manipulation arguments.[16] On a stability-centric view, the boss would be the causal source of the murder while the assassin would not. This would mean that though the assassin could not be morally responsible for the murder, the boss could be. At first pass, the stability-centric view seems less objectionable than the reliability-centric view. Perhaps powerful bosses who exert their influence to get their employees to do their bidding *are* the causal sources of their employees' actions and should be held solely morally responsible for them.

But problems arise for the stability-centric version of DN's account as well. Sometimes the most stable causal relation fails to deliver a morally responsible agent. For instance, suppose there is a car accident (C=c). There are two causal variables: the presence of a pothole (P=p) and a driver who is driving in a culpably distracted manner (D=d). Both the pothole and the driver driving distractedly are actual causes of the crash—if the driver hadn't been distracted, the crash would not have occurred (holding fixed the actual background conditions) and likewise if the pothole wasn't there (again, holding fixed the actual background conditions).

[15] In general, reliability indicates the limits of a causal relationship. Consider, for example, Hooke's law: F = -kx, where x is the elongation of a spring, k is a constant characteristic of the spring, and F is the restoring force exerted. Hooke's law expresses a genuine causal relation between elongation and force, but nevertheless breaks down at a certain value of x—one can only stretch a spring so far. By contrast, it is often supposed that the laws of (ideal) fundamental physics should be perfectly reliable (Woodward 2007: 76–80).

[16] Indeed, Usher (2018), who defends his own stability-centric view of causation, characterizes DN's position as only concerned with stability (footnote 14).

It's possible that the causal relation P → C is more stable than D → C, i.e. the presence of the pothole predicts the crash under a wider range of background conditions than does the driver being distracted. If this is the case, P would be the causal source of the crash and not the driver, rendering the driver an inappropriate target for responsibility. But surely the driver is morally responsible for the crash. His driving distractedly was an actual cause of the crash—if D had taken a different value, the crash would not have occurred (holding fixed the background conditions).

Furthermore, even if we limit ourselves to cases involving only *agential* variables that are stably related to the effect, the problem remains. Suppose that a large truck was also involved in the accident. Its driver was alert and driving safely near the pothole and caused the crash to be far worse than it would have been otherwise ($T=t \rightarrow C=c$). It's possible that the relation between the truck driver's behavior and the accident is more stable than D → C. Perhaps given the momentum of the truck, the accident was bound to occur under a wider range of road conditions, other drivers' behavior, weather, etc. DN's account would identify the truck driver's behavior as the causal source of the crash, and the distracted drive would be off the hook. But again, it seems that the distracted driver is morally responsible in this case as well. The truck driver, though her behavior relates more stably to the crash, did nothing wrong. But the distracted driver did do something wrong, even though his behavior relates less stably to the crash.

Stability, like reliability, can tell us many things about the causal relationships between events. For instance, knowing that the presence of a pothole (or the truck driver's behavior) relates more stably to the crash than any other causal variable tells us, other things being equal, that the most effective way to prevent the crash would have been to fill in the pothole (or change the truck driver's behavior). But the connection between stability and moral responsibility is far from straightforward. An agent can be morally responsible for an outcome without relating to it in the most stable manner, and the most stable causal sources are often not the appropriate targets of praise, blame, and moral responsibility.[17]

4.d. Go scalar

Up until this point, we've been conceiving of DN's account of causal sourcehood and moral responsibility in absolute terms; an agent can only be morally responsible for an outcome if she is its causal source, and an agent can only be the causal source of an outcome if she bears the strongest invariance relation to it. But perhaps DN could weaken either of these constraints to make their account more flexible.

First, DN could argue that causal sourcehood is not a necessary condition for moral responsibility, though it is relevant. In fact, DN suggest such a view later in their paper:

> 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…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…) outside their CAS. (DN 2017: 1272)

On this view, moral responsibility comes in degrees and agents whose CASs aren't the causal sources of their actions can still be morally responsible for them, but to a mitigated degree.[18]

[17] The above reasoning counts against Usher's (2018) stability-centric account of responsibility as well. Usher argues that causally responsibly for an event depends on an agent's beliefs and intentions relating stably to the event in question. Usher goes on to argue that if an agent possesses the relevant CAS, then the degree to which an agent is causally responsible determines the degree to which she is morally responsible as well. But, as we've seen above, the relationship between stability and moral responsibility is not this straightforward. An agent can be morally responsible to a high degree for an action even if her intentions and beliefs don't relate to it in a particularly stable way. Furthermore, another agent, who possesses the relevant CAS, can be not at all morally responsible for an event even though she relates to it more stably than any other causal variable.

[18] Developing an account of causal sourcehood that can make sense of moral responsibility as coming in degrees may prove to be a promising strategy. Historically, theories of moral responsibility have focused almost exclusively on the threshold of moral responsibility, but this is changing (e.g., Coates and Swenson 2013; Nelkin 2016; XXX). Usher (2018) also takes it to be a significant virtue of his interventionism-inspired view of teleological control that it makes room for degrees of responsibility. And, those developing models of causal responsibility in computer science and cognitive science have also focused on modeling degrees of responsibility (e.g., Chockler and Halpern 2004; Lagnado et al. 2014). However, these latter accounts are careful

It's difficult to assess this view because it remains unclear how and why the presence of an alternative causal source mitigates, but does not eliminate, moral responsibility. But, setting this worry aside for now, the view also generates counterintuitive verdicts with respect to the cases considered above. For instance, consider a version of the pothole case in which the pothole is deemed the causal source of the crash.[19] On this version of DN's view, the distracted driver could be morally responsible for the crash, but to a lesser degree due to the existence of an alternative causal source. But it strikes us as counterintuitive that the presence of a pothole should mitigate the degree to which the distracted driver is responsible *at all*. The driver's being distracted is an actual cause of the crash—if he hadn't been distracted, the crash would not have occurred (holding fixed the actual background circumstances). Culpably distracted drivers often crash into things. Surely the mere existence of the objects into which they crash doesn't mitigate the degree to which these drivers are morally responsible. Of course, if the pothole would have caused the crash even if the driver wasn't distracted (holding fixed the actual background conditions), this would plausibly mitigate the degree to which the driver is responsible. But this would mean that the driver's being distracted is not an actual cause of the crash, which is contrary to our supposition.[20] Perhaps these counterintuitive consequences would be tolerable if there existed independent motivation for the notion that the failure of an agent to be the causal source of their action mitigates but does not eliminate moral responsibility. But DN offer no such independent motivation, and thus, there is little to recommend the view.

to distinguish between causal responsibility and moral responsibility. In fact, Chockler and Halpern argue that agents can be blameworthy without being causally responsible (2004: 95).

[19] There are two ways this could go. First, one could adopt a stability-centric view of causal sourcehood, which would have as a consequence that the pothole is the causal source of the crash, for the reasons discussed above. Second, one could argue that the pothole is both most stably and most reliably related to the crash, and thus unambiguously the causal source. Perhaps the physical dimensions and the location of the pothole under a wide range of values would still predict the occurrence of the crash, in which case the causal relation involving it has a good claim to being the most reliable.

[20] The same can be said, mutatis mutandis, for a version of the truck driver case in which the truck driver's behavior is deemed the causal source of the crash.

Rather than deny that causal sourcehood is necessary for moral responsibility, DN could make their view more flexible by arguing that causal sourcehood (and hence, moral responsibility) can come in degrees. On this view, causal sourcehood can be achieved even if a cause doesn't bear the strongest invariance relation to an outcome and there can be multiple causal sources of a single outcome. One way of developing this view would be to argue that the degree to which an agent is the causal source of an outcome is determined by the strength of the overall causal invariance relation (i.e., some function of stability and reliability). In fact, DN suggest such a view in footnote 8 (2017: 1263) and footnote 13 (2017: 1268–1269) of their paper.

On its face, this view seems quite plausible. Consider a version of the boss and assassin case in which the boss bears a stronger overall invariance relation to the murder than the assassin.[21] On this version of a scalar view, one could argue that both the boss and the assassin are causal sources of the murder but to different degrees—the boss is more of a causal source than the assassin. So, both the boss and the assassin could be morally responsible for the murder, though the boss would be morally responsible for the murder to a greater degree. There seems to be nothing particularly counterintuitive about this result, though intuitions certainly could vary. But the real problem with this view is that it undermines DN's soft-line response to manipulation arguments.

On this scalar view of causal sourcehood, both Danny's and Manny's decision to steal the money would each qualify as a causal source of their stealing the money. And, because their decisions fulfill the stability and reliability conditions on causal sourcehood to exactly the same degree, Manny and Danny would be causal sources to the same extent. So, Manny and Danny would be equally

[21] Perhaps the boss meets the stability condition to a greater degree than the assassin meets the reliability condition or perhaps they each meet their respective conditions to the same degree, but stability has a greater weight in determining overall invariance.

morally responsible for stealing the money. Thus, there would no longer be a relevant difference between Manny and Danny, and DN's objection to NoDif would be undermined.[22]

However, there is one difference between the cases of Manny and Danny on this view: While Danny and Manny are equally morally responsible for stealing the money, there is someone who is *more* responsible than Manny for his stealing the money—Diana—which is not the case for Danny. However, the presence of Diana doesn't affect the degree to which Manny is a causal source of, and thus morally responsible for, stealing the money. On the present approach, the degree to which Manny's decision is a causal source of the stealing does not depend on the degree to which Diana is a causal source of the stealing and the extent to which Manny is morally responsible has nothing to do with the extent to which Diana is morally responsible. The two are simply independent of one another.[23]

5. What about Context?

In this paper, we've argued that there is likely no satisfying way to resolve the conflict between stability and reliability on DN's account of causal sourcehood. But perhaps the original case that highlighted the difference between these two distinct notions of invariance—the boss and assassin case—is flawed. One could argue that we haven't evaluated the invariance relations between the boss's decision (B), the assassin's decision (A), and the murder (M) correctly. For example, we argued that, when evaluating the stability of the relationship between A and M, we must ignore any causal contribution

[22] If one finds this scalar approach to causal sourcehood and responsibility promising, one could adopt it to develop a hard-line response to manipulation arguments.

[23] One could stipulate a kind of "countable additivity" requirement that degrees of causal sourcehood must sum to 1. If this were the case, then the presence of Diana would make Manny considerably less responsible, as she would take up much of the sourcehood, and hence, responsibility. This claim runs contrary to much of the philosophical literature on responsibility. For example, Zimmerman (1985) argues that multiple agents can be fully responsible for the outcomes of group actions. Additionally, experimental work indicates that the folk do not comply with a countable additivity requirement when they assess responsibility (Lagnado et al. 2014). Finally, such a principle would make much of DN's argument unnecessary. The mere presence of a novel cause, with even the most modest degree of sourcehood, would serve to differentiate Manny from Danny. This fails to capture the key feature of the case—that Manny isn't the source of his stealing the money.

from B. But one might argue that the assassin's CAS encompasses the fact that her boss hired her to kill his enemy. After all, many accounts of free will rely on notions of reasons-responsiveness (Fischer & Ravizza 1998) and reasons sensitivity (Sartorio 2016), so it would make sense to include the reason to engage in the blameworthy behavior as a feature of the agent's compatibilist agential structure. Other accounts rely on first and second-order desires (Frankfurt 1971), and perhaps the boss's request causes the assassin to develop a first-order desire to kill the enemy that is in-line with a second-order desire to desire things that will make one's boss happy (or at least not angry). And if this is the case, then even when we perform an intervention on A and ignore the role of B, we still must take into account that the boss asked the assassin to kill the enemy (either as a reason for acting or a desire to be fulfilled), in which case A would bear at least as stable a relation to M as B would.

However, it's not clear what this particular amendment will do to safeguard DN's account of causal sourcehood. Because the relationship between A and M is arguably more reliable than the relationship between B and M, A would meet both the stability and reliability conditions of DN's account of causal sourcehood, and would thus be the causal source of M. But this would mean that while the assassin could be held responsible for the murder, the boss would not be at all responsible (or responsible to a lesser degree on a scalar version of DN's account). But this is the same counterintuitive result that led us to reject the reliability-centric account of causal sourcehood.

Of course, it may be possible to generate intuitive results by using views of free will and moral responsibility to arrange the variables that represent the boss's and assassin's decisions in a causal model. But there is something suspicious about relying on views of free will to determine how to construct our causal models. DN originally looked to interventionist approaches to causation to settle the question about whether Manny was free to, and morally responsible for, stealing the money. There is something circular, or at least uninformative, about using views of free will and moral responsibility to inform features of an interventionist causal model when the questions we're trying to answer are about free will and moral responsibility. This isn't to say that it's illegitimate to rely on such views in constructing an interventionist model when the questions we're trying to answer are about something other than free will. Nor do we wish to argue that there are *no* informative or illuminating uses of interventionism within the free will debate in general. In the next and final section of this paper, we briefly gesture at one possible application of DN's account of causal sourcehood.

6. Conclusion

DN's account of causal sourcehood provides effective strategies for intervening on the world (cf., Cartwright 1979). But the relationship between moral responsibility and intervention strategies is not straightforward. In a traffic accident, for instance, we may wish to know the best way to prevent such a thing from occurring, i.e., we may want to know which variable we should intervene on to change the outcome. But, the most effective point of intervention needn't be a responsible agent. It could just as well be a pothole or a truck driver safely doing her job. And, importantly, agents who do not qualify as the most effective points of intervention can still act freely and be held morally responsible. Thus, we cannot hope to read off facts about moral responsibility from facts about effective intervention strategies.

Though DN's account should remain silent on backwards-looking questions about praise, blame, and moral responsibility, it can still play an important role in discussions of forward-looking questions about these practices. For example, the determination of effective strategies can be quite useful when we want to evaluate our practices of praise and blame. These practices can have significant and unintended effects on their targets. By using an interventionist model, we could evaluate the causal relationships between instances of praise and blame and their effects on an agent's CAS and/or future behavior. This would allow us to determine the most effective practices of praise and blame for agents' moral and behavioral reform. Such information would not only be useful in evaluating the effectiveness of individual instances of praise and blame, but it could also inform theories of the morality of praise and blame more generally.

References

Campbell, John (2010) "Control Variables and Mental Causation." *Proceedings of the Aristotelian Society*. 110, pp. 15–30.

Cartwright, Nancy (1979) "Causal Laws and Effective Strategies." *Noûs* 13. 4, pp. 419–437.

Coates, Justin and Swenson, Philip (2013) "Reasons-Responsiveness and Degrees of Responsibility." *Philosophical Studies* 165. 2, pp. 629–645.

Chockler, Hana and Joseph Halpern (2004) "Responsibility and Blame: A Structural-Model Approach," *Journal of Artificial Intelligence Research* 22, pp. 93–115.

Deery, Oisín and Nahmias, Eddy (2017) "Defeating Manipulation Arguments: Interventionist Causation and Compatibilist Sourcehood." *Philosophical Studies* 174. 5, pp. 1255–1276.

Fischer, John M. (2000) "Responsibility, History, and Manipulation." *Journal of Ethics* 4. 4, 385–391.

Fischer, John Martin and Ravizza, Mark (1998) *Responsibility and Control*. Cambridge: Cambridge University Press.

Frankfurt, Harry (1971) "Freedom of the Will and the Concept of a Person." *Journal of Philosophy*. 68. 1, pp. 5–20.

Ismael, J. T. (2013) "Causation, Free Will, and Naturalism." In Kincaid, A., Ladyman, J., and Ross, D. (eds.) *Scientific Metaphysics*. New York: Oxford University Press, pp. 208–235.

Ismael, J. T. (2016) *Why Physics Makes Us Free*. New York: Oxford University Press.

Lagnado, David, Gerstenberg, Tobias, and Zultan, Ro'i (2014) "Causal Responsibility and Counterfactuals." *Cognitive Science* 37. 6, pp. 1036–1073.

McKenna, Michael (2008) "A Hard-line Reply to Pereboom's Four-case Argument."

Philosophy and Phenomenological Research. 77. 1, pp. 142–159.

Mele, Alfred (2013) "Manipulation, Moral Responsibility, and Bullet Biting." *Journal of Ethics*. 17. 3, pp. 167–184.

Mele, Alfred (2005) "A Critique of Pereboom's 'Four-Case' Argument for Incompatibilism." *Analysis* 65, pp. 75–80.

Nelkin, Dana (2016) Difficulty and Degrees of Moral Praiseworthiness and Blameworthiness, *Noûs* 50. 2, pp. 356–378.

Pearl, Judea (2009) *Causality*. Cambridge: Cambridge University Press.

Pereboom, Derk (2014) *Free Will, Agency, and Meaning in Life*. New York: Oxford University Press.

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

Roskies, Adina (2012) "Don't Panic: Self-authorship Without Obscure Metaphysics." *Philosophical Perspectives*. 26. 1, pp. 323–342.

Usher, Marius (2018) "Agency, Teleological Control and Reliable Causation." *Philosophy and Phenomenological Research*. https://doi.org/10.1111/phpr.12537

Wolf, Susan (1990) *Freedom Within Reason*. New York, NY: Oxford University Press.

Woodward, James (2003) *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

Woodward, James (2006) "Sensitive and Insensitive Causation." *Philosophical Review* 115. 1, pp. 1–50.

Woodward, James (2007) "Causation with a Human Face." In Price, A. & Corry, R. (eds.). *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford: Oxford University Press.

Zimmerman, Michael (1985) "Sharing Responsibility." *American Philosophical Quarterly* 22. 2, pp. 115–122.