The Causal and the Moral

by

Ana Carolina Sartorio

B.A. Universidad de Buenos Aires, 1996

Submitted to the Department of Linguistics and Philosophy
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology
[June 2003]
May 2003

Author.....
Department of Linguistics and Philosophy

Certified by.....
Stephen Yablo
Professor of Philosophy
Thesis Supervisor

Accepted by.....
Vann McGee
Chair of the Committee on Graduate Studies

Abstract

My dissertation is about the following two questions:

The causal question: When is something a cause of something else?
The moral question: When is someone morally responsible for something?

I examine the way in which these questions overlap. I argue that, in some important respects, the relation between the causal and the moral question is *tighter* than people have taken it to be, but, in other important respects, it is *looser* than people have taken it to be.

The dissertation consists of three chapters. Each of the chapters is a self-contained paper, but the three papers are interconnected in various ways. Chapters 1 and 2 are concerned with how the causal question and the moral question intersect, and Chapter 3 is concerned with how they come apart.

In Chapter 1, I lay out a view of causation according to which causing is a particular way of making a difference. I show that an advantage of this view is that it carves up a concept of cause that is particularly well suited for the work causation does in moral theory.

In Chapter 2, I argue that a moral asymmetry that exists between actions and omissions has a causal basis. I argue that the conditions under which actions and omissions make us morally responsible are different, and that this is so because the causal powers of actions and omissions are different.

In Chapter 3, I argue against the received view about the relation between causation and moral responsibility, according to which being responsible for something requires causing it. I offer an alternative picture according to which causation is a necessary condition for the *transmission* of responsibility, although not for the *existence* of responsibility itself.

Thesis Supervisor: Stephen Yablo
Title: Professor of Philosophy

Acknowledgments

I would like to specially thank my advisor, Steve Yablo. Steve was a great inspiration. He is one of the most imaginative, enthusiastic, and charismatic persons I know, and it is remarkably easy for him to transmit that enthusiasm to other people. His approach to philosophy changed my view of philosophy. I benefited enormously from my talks with him, and he helped me a great deal in writing this dissertation, not just in putting all the pieces together, but also in developing the main ideas behind it.

I also worked with Ned Hall and Judy Thomson. Ned has been spectacularly generous with his time and support. He is also one of the sharpest people I know. Nothing gets past him: if a view has counterexamples, he'll most likely detect them immediately. Judy has been a great influence on me. Without hesitation, I can say that, had it not been for Judy, I would have never become interested in ethics. I am particularly grateful to her for teaching me how to think about metaphysical and moral issues in connection with each other. She also has a way of doing philosophy that I admire, one in which there is a delicate combination of clarity, rigor, and deep philosophical ideas.

The MIT philosophy program is famous for its unique atmosphere. I benefited enormously from it. I made a lot of good friends and I learned much from many different people. I am particularly grateful to Sarah McGrath, who read several versions of each of the chapters and made many helpful suggestions. Special thanks also to Tyler Doggett, Andy Egan, Liz Harman, Jim John, and Agustín Rayo. Many thanks, too, to the members of the MATTI graduate student reading group at MIT and the Harvard/MIT Friends and Eminees reading group for much helpful discussion.

Back home, thanks to Eduardo Barrio, Eduardo Flichman, Eleonora Orlando and Federico Penelas. For their very generous financial support, I am grateful to the MIT Department of Linguistics and Philosophy and to Fundación Antorchas in Argentina.

Finally, for their encouragement and support, thanks to my family and friends, who helped to make my stay at MIT such an enjoyable experience. In particular, thanks to my husband, Juan Comesaña,

who helped me in countless different ways. He read more versions of the whole manuscript than anyone else and made many invaluable suggestions. He also put up with a tediously long commute between Cambridge and Providence during our five years of graduate school, so that I didn't have to do it. And he enriched my life in graduate school and outside of graduate school in every way possible. Some people claim that it is better to marry someone outside your field to lead a richer life as a person. I think that they are terribly wrong.

Contents

Chapter 1

Causes As Difference-Makers

1. Introduction

David Lewis wrote in "Causation:"

> We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it.[1]

Call this idea, according to which a cause is a "difference-maker," *the difference-making idea.* The difference-making idea famously motivated Lewis's counterfactual theory, an attempt to analyze the concept of cause in terms of the relation of counterfactual dependence between events.[2]

However, as we will see shortly, the counterfactual theory ends up misrepresenting the difference-making idea: it counts as causes things that aren't difference-makers. We should then look for an alternative way of spelling out the difference-making idea. This is what I will do in this paper. I will make a new proposal on how causes are difference-makers, and I will argue that the new proposal succeeds in capturing the difference-making idea.

Two words of clarification are in order. First, the view that I will defend here is not an *analysis* of causation. It sets a constraint on the concept of cause, and thus it helps to carve up the concept, while at the same time leaving some room for different ways of pinning it down. Second, this paper is an attempt to establish how *best* to capture the difference-making idea; it is not—at least, not primarily—a defense of the claim that we should endorse it. However, at the end of the paper I will point to an

---

[1] Lewis (1986a), pp. 160-1.

[2] In recent years, Lewis's theory gave rise to an array of revisions and adjustments, all of which attempt to analyze the concept of causation, ultimately, in terms of counterfactual dependence between events. Two examples are McDermott (1995) and Lewis himself in his later work, Lewis (2000).

important advantage of endorsing the difference-making idea: I will argue that a concept of cause that results from embracing it is particularly well suited for the work causation does in moral theory.

2. Two ways of making a difference

I will start by illustrating how the difference-making idea motivated Lewis's counterfactual theory and how, despite this fact, Lewis's theory failed to capture it. This discussion will then serve to motivate my own proposal.

On its first pass, Lewis's counterfactual theory (CT) states that a cause is something without which the effect wouldn't have occurred:

> CT (First Pass): C causes E if and only if E counterfactually depends on C, i.e., if C hadn't happened, then E wouldn't have happened.

That is, a cause makes a difference to its effect in that the effect wouldn't have occurred without the cause. This first pass has obvious counterexamples, such as the following:

> *Assassination*: Assassin shoots Victim and, as a result, Victim dies. However, Backup is waiting in reserve. Had Assassin not shot, Backup would have, and Victim would still have died (in a very similar way, at around the same time, etc.).

Victim's death (intuitively, the *same* death) would still have occurred if Assassin hadn't shot; hence, CT (First Pass) entails that Assassin's shooting wasn't a cause of Victim's death. But, clearly, it was.

In order to get around this problem, Lewis takes the *ancestral* of counterfactual dependence:

> CT (Second Pass): C causes E if and only if there is a chain of stepwise counterfactual dependence from C to E.[3]

In Assassination, there is a chain of stepwise counterfactual dependence from Assassin's shooting to Victim's death, via the intermediate event of Assassin's bullet's heading towards Victim. Assassin's

[3] Lewis (1986a), p. 167.

bullet heading towards Victim depends on Assassin's shooting, for, had Assassin not shot, Assassin's bullet wouldn't have been heading towards Victim. In turn, given that Backup didn't shoot,[4] Victim's death depends on Assassin's bullet's heading towards Victim, for, had Assassin's bullet not been heading towards Victim, Victim wouldn't have died. Hence, CT (Second Pass) yields the right result: Assassin's shooting caused Victim's death.

Interestingly, however, this second pass has an important drawback. By letting chains of dependence in, Lewis counts too many things as causes, including things that, intuitively, don't make a difference to the effect. Hence the move from the first pass to the second pass is a step away from the difference-making idea, which originally served to motivate the theory. Consider, for instance, the following case:

> *Switch*: Victim is stuck on the railroad tracks. A runaway train is hurtling down the tracks when it approaches a switch. I flip the switch, and the train turns onto a side track. However, the tracks reconverge a bit further ahead, before the place where Victim is standing. Victim dies.

Is my flipping the switch a cause of Victim's death? According to CT (Second Pass), it is, for there is a chain of stepwise counterfactual dependence from my flipping the switch to the death, via the intermediate event of the train running on the side track. This emerges as follows: The train's running on the side track depends on my flipping the switch, for, had I not flipped the switch, the train wouldn't have been running on the side track. In turn, given that the train switched tracks and thus it is no longer on the main track, Victim's death depends on the train's running on the side track. For, if it hadn't been running on the side track, then, given that it is not running on the main track, the train would not have reached Victim and killed him.[5] Hence, the train's running on the side track depends on my flipping the switch,

---

[4] According to Lewis, the standard contexts of evaluation of counterfactuals are not "backtracking". Thus, in considering a counterfactual of the form "If C hadn't occurred, E wouldn't have occurred," we must hold fixed as much of what happened before C as possible. See Lewis (1986b).

[5] Where would have the train gone, then? This depends on the details of our theory of counterfactuals. Maybe it would have derailed, or it would have miraculously vanished. In any case, it is clear that it wouldn't have reached Victim, given that Victim could only be reached via one of the tracks.

and Victim's death depends, in turn, on the train's running on the side track. As a result, CT (Second Pass) entails that my flipping the switch caused Victim's death in Switch. However, this result clashes with the difference-making idea: intuitively, my flipping the switch did not make a difference to Victim's death. To be sure, it made *some* difference, e.g., it made the death happen via the train's running on the side track. But, intuitively, this is a difference that does not matter causally.[6]

It might be objected that our causal intuitions about Switch are morally tainted: that our intuitive verdict about Switch arises from a confusion between what I am causally responsible for and what I am morally responsible for. Clearly, I am not morally responsible for the death in virtue of having flipped the switch. Thus, since it is hard to keep the moral intuition apart from the causal intuition, it is natural to think that I don't cause the death by flipping the switch. But I might cause it without being morally responsible for it; after all, we cause many things that we are not morally responsible for.

However, we can address this worry by imagining a variant of the case where no moral agents are involved. Imagine, for instance, that what flipped the switch is a gust of wind, and what was lying on the track up ahead is an apple. This doesn't change our causal intuitions: it still seems that the flipping of the switch isn't a cause of the outcome—in this case, the squashing of the apple. This suggests that our intuition that my flipping the switch didn't cause the death in Switch is genuinely causal, not merely moral.[7]

We have seen that CT does not succeed in fully capturing the difference-making idea: its second pass (to which one can be naturally driven upon realizing that the first pass has clear counterexamples) counts as causes things that, intuitively, don't make a difference to the ensuing outcomes. In particular,

---

[6] Lewis's most recent attempt, the "causation as influence" view (in Lewis (2000)), has the same kind of problem. For there Lewis analyzes causation as the ancestral of the influence relation. As a result, he counts as causes things that, intuitively, make no difference to the effects.

[7] For attempts to rescue the intuition that my flipping the switch isn't a cause in cases of this sort, see Rowe (1989), and Yablo (2002) and (forthcoming). I do not wish to suggest, however, that these authors would agree with the proposal I will offer shortly.

we have seen that, in Switch, the flip doesn't seem to make a difference to Victim's death. However, CT counts the flip as a cause of the death.

So we should look for a new way to capture the difference-making idea. I propose that we start by focusing on Switch, where CT failed. One thing that catches the eye about Switch is that, just as the *flip* doesn't make a difference to the death, the *failure to flip* wouldn't have made a difference to the death either. In other words, whether or not I flip the switch doesn't make a difference to the death; it only helps to determine the route that the train takes before reaching Victim. This suggests that what might be missing in Switch is some kind of *asymmetry* between my flipping the switch and my failing to flip the switch. Maybe the reason that my flipping the switch doesn't make a difference is that the contribution that it makes is not more important than the contribution that its absence would have made. Maybe, for something to be a cause, it must make a contribution that somehow *outweighs* the contribution that its absence would have made.

How could we make this thought more precise? Here is a natural suggestion. Causes are difference-makers in that the following principle, the *Causes as Difference-Makers* principle, is true:

CDM: If C caused E, then the absence of C wouldn't have caused E.

According to CDM, a cause contributes more to the effect than its absence would have contributed to it in that the absence of the cause wouldn't have been a cause itself.[8]

Consider what CDM would say about Switch. I have pointed out that, intuitively, the contributions that the flip made and that the failure to flip would have made are on a par. Hence, it is likely that, were we to count the flip as a cause, we would also have to count the failure to flip as a cause.[9] But CDM doesn't allow this. So, CDM would entail that the flip *isn't* a cause. As a result, CDM would help explain our reluctance to count the flip as a cause of the death in Switch.

---

[8] I intend this to apply to both "positive" and "negative" causes. For instance, C could be an omission, in which case the absence of C would be an action (more on this below).

[9] On the assumption that there is causation by omission. I discuss this assumption shortly.

I will argue that CDM succeeds in capturing the difference-making idea. The rest of this paper is concerned with clarifying the content of CDM, arguing for its truth, and examining its most important consequences. First, however, let me briefly compare CDM with CT.

An important difference between CDM and CT is that, unlike CT, CDM cannot be regarded as a reductive analysis of causation, i.e. as an analysis of the concept of cause in purely non-causal terms. CDM is, rather, a *constraint* on theories of causation: a condition that the true analysis of causation (if there is such a thing) would have to meet. Another important difference between CDM and CT is in the way each attempts to capture the difference-making idea. We have seen that CT attempts to cash out the difference-making idea in terms of the relation of counterfactual dependence between events. According to CT, a cause makes a difference in that, if it hadn't occurred, then some event intimately related to the effect wouldn't have occurred. (On the first pass, the effect itself wouldn't have occurred; on the second pass, some event in the chain of events leading to the effect wouldn't have occurred.) According to CDM, a cause makes a difference by determining, not the *events* that occur in the actual and counterfactual scenarios, but the *causal relations* that obtain in the actual and counterfactual scenarios: whether a cause occurs makes a difference to whether there is a causal relation linking an event or its absence (according as the event is present or absent) to the effect.

3. Events, absences, and the stringency of CDM's demands

In this section I explain the content of CDM in more detail and I illustrate with examples. In the next two sections I lay out my argument for CDM.

I will be assuming that absences can be causes, or, as it is sometimes put, that there is *causation by omission*. This is a reasonable assumption. Intuition dictates that there is causation by omission (intuitively, the absence of rain can cause a drought, and a mother's failure to feed her child can cause the child's death), and the majority of philosophers have followed intuition on this score. This is not to say that the assumption that there is causation by omission is trouble-free; there are problems generated by letting omissions be causes, but saying that causation by omission is impossible still seems like an

overreaction to such problems.[10] The assumption that there is causation by omission prevents CDM from being *trivially* true. If there were no causation by omission, then, clearly, it could never be the case that both an event and its absence would have caused the same effect in the scenarios where they obtain, simply because an event's absence could never cause anything.

With this assumption in place, let us examine in more detail what CDM says. On the one hand, CDM makes a claim about how the causal powers of *events* constrain the causal powers of the corresponding *absences*. Suppose that I write a letter to my mother and that makes her happy. Then CDM claims that, given that my writing her a letter caused her to be happy, my failure to write to her wouldn't have caused her to be happy. This is to say, had I not written a letter to my mother, my failure to write to her wouldn't have caused her to be happy.

Some words of clarification are in order. First, what is a failure? In particular, what is *my failure to write a letter to my mother*? I will adopt a common convention according to which a failure is the failure of any event of a certain type to occur.[11] On this view, the failure to write a letter to my mother obtains just in case no event of a certain type—a writing a letter to my mother by me, at a certain time, or within a certain interval of time—occurs. More generally, if C is an event, then the absence of C obtains just in case no C-type event occurs. In a case of this sort, where C is an event, CDM claims that, if C caused E, then, had no C-type event occurred, the failure of a C-type event to occur wouldn't have caused E.

Second, how should we understand the counterfactual claims that CDM makes? In particular, how should we understand the claim: "Given that C caused E, *had no C-type event occurred, then the failure of a C-type event to occur wouldn't have caused E*"? We can interpret it in the standard way, i.e.

---

[10] Causation by omission would be a problem if, for instance, one believed that the causal relata are events. For, on many views, omissions aren't events. On the other hand, it is important to allow for causation by omission in order to preserve the important connection that seems to exist between causation and moral responsibility. If a mother doesn't feed her baby, it seems that she is morally responsible for the baby's death in virtue of having caused his death by not feeding him. But this is causation by omission. For discussion of causation by omission, see Dowe (2001), McGrath (ms) and Thomson (2003).

[11] See, e.g., Lewis (1986a), p. 189.

by appeal to possible-worlds semantics. Take the closest possible world where no C-type event occurs; that is a world where the failure of a C-type event to occur obtains. What CDM says is that the failure of a C-type event to occur doesn't cause E in that world, given that C caused E in the actual world.

Now, CDM also makes a claim about how the causal powers of *absences* constrain the causal powers of certain specific *events*. Suppose that I fail to phone Grandma on her birthday and this makes her sad. Then CDM claims that, given that my failure to phone Grandma on her birthday caused her to be sad, had I phoned her on her birthday, my phoning her wouldn't have caused her to be sad. Again, we can interpret this in terms of possible worlds. Take the closest possible world where I phone Grandma. CDM says that, in that world, my phoning Grandma doesn't cause her to be sad. More generally, if C is *the failure of an event of a certain type to occur,* i.e. an absence, then CDM claims that, if C caused E, then, had an event of the relevant type been present, it wouldn't have caused E. This is to say, in the closest possible world where an event of that type occurs, that event doesn't cause E, given that the absence of an event of that type caused E in the actual world.

I have explained the content of CDM in some detail. Now I will discuss its force. CDM imposes a constraint on theories of causation. How hard is it for a theory of causation to comply with CDM?

The most interesting and controversial claim that CDM makes concerns outcomes that would still have occurred in the absence of the cause, i.e., cases where the outcome doesn't counterfactually depend on the cause. In cases where the outcome counterfactually depends on the cause, CDM is met in a straightforward way. To see this, imagine that Assassin shoots and, as a result, Victim dies. However, had Assassin failed to shoot, Victim would have lived. Then it is trivially true that, whereas Assassin's shooting caused Victim's death, Assassin's failing to shoot wouldn't have caused Victim's death: it wouldn't have caused the death because the death wouldn't have occurred if Assassin had failed to shoot.

So, in order to measure the stringency of CDM's demands, we must focus on cases where the outcome doesn't counterfactually depend on the cause. The question then becomes: How hard is it to meet CDM's demands in *those* cases? I will show that it is quite hard. As a matter of fact, coming up

with a theory of causation that complies with CDM is no easy task.[12] By way of example, I will briefly review two different types of theories of causation and I will show that they both fail to meet CDM.

A first type of theory that clashes with CDM is a type of theory according to which helping to determine the causal route to an effect is sufficient for causing the effect.[13] When I flip the switch in Switch, my flipping the switch makes the train run on the side track before it reaches Victim; hence, it contributes to determining the causal route to Victim's death. So a theory of the type we are envisaging entails that my flipping the switch causes the death. However, had I *failed* to flip the switch, my failure to flip the switch would also have contributed to determining the causal route to the death, for it would have made the train run on the *main* track before it reached Victim. So a theory of this type would also entail that, had I failed to flip the switch, my failure to flip the switch would have been a cause of the death. In other words, according to a theory of this type, no matter what I did in Switch, I would have caused the death. This contradicts CDM.

A second type of theory of causation that fails to comply with CDM is a classical "regularity" view, such as Mackie's view.[14] According to Mackie, C is a cause of E just in case there is a set of occurring conditions containing C that, when conjoined with some lawful regularity, entails E, and that doesn't entail E when C is removed from the set. Consider what this view would say about Assassination. In Assassination, given that Assassin shot, there is a set of occurring conditions containing the fact that Assassin shot that, when conjoined with some lawful regularity, entails the fact that Victim died, but that doesn't entail this when the fact that Assassin shot is removed from the set. This set of conditions includes, for instance, the fact that Assassin's gun had working bullets, the fact that it was

---

[12] A theory that clearly meets CDM is CT (First Pass). On this view, if C is a cause of E, then E wouldn't have occurred in C's absence; hence, it is clear that the absence of C wouldn't have caused E. However, as we have seen, there are clear counterexamples to this view. As a result, no one seems to hold it.

[13] An example of a theory of this type is CT (Second Pass).

[14] Mackie (1993). See, in particular, the example on p. 43. There Mackie seems to suggest that his view has the consequence that it is possible for the fact that an event occurs to cause an outcome when the fact that the event doesn't occur would also have caused it. (Notice that, for Mackie, the causal relata are facts, not events. But nothing essential hangs on this.)

aimed at Victim, etc. Hence, Mackie's view would say that Assassin's shooting was a cause of Victim's death. But now imagine that Assassin hadn't shot, in which case Backup would have shot. Then there would have been a set of occurring conditions containing the fact that Assassin *didn't* shoot that, when conjoined with some lawful regularity, entails the fact that Victim died, but that doesn't entail this when the fact that Assassin didn't shoot is removed from the set.[15] This set of conditions includes, for instance, the fact that Backup intended to shoot just in case Assassin didn't shoot, the fact that Backup's gun had working bullets, etc. Hence, Mackie's view would say that Assassin's shooting caused the death but, had Assassin not shot, his failure to shoot would also have caused the death. This contradicts CDM.

We have seen that CDM imposes a highly demanding constraint on theories of causation; by way of example, I have shown that two importantly different types of theories of causation fail to comply with it. In what follows, I argue that CDM succeeds in capturing the difference-making idea. Hence, if we are to respect the difference-making idea, we should reject any theory of causation that fails to comply with CDM.

4. Argument for CDM (Part I)

I will argue for CDM by showing that the best candidate counterexamples to CDM are not genuine counterexamples. I will look at two paradigm cases where it is most plausible to think that both an event and its absence would have caused an outcome, and I will argue that they fail. Since they fail, and since they are the best attempts at counterexamples, I will conclude that there is good reason to believe that CDM is true.

For ease of exposition, I will focus on the specific claim that CDM makes about actions and omissions of agents, but the argument is intended to have full generality. When restricted to actions and omissions of agents, CDM reads:

---

[15] This last part is true because Backup wouldn't have shot *unless* he saw that Assassin didn't shoot.

CDM (A/O): If an agent's acting in a certain way caused E, then, had the agent failed to act that way, the agent's failing to act that way wouldn't have caused E and, vice-versa, if an agent's failing to act in a certain way caused E, then, had the agent acted that way, the agent's acting that way wouldn't have caused E.

A counterexample to CDM (A/O) would have to be a case where, in the scenario where the agent acts in the relevant way, the agent's action causes an outcome E and, also, in the scenario where the agent doesn't act in the relevant way, the agent's omission causes E.

Let us single out, in particular, the following three desiderata that a counterexample to CDM (A/O) would have to meet. First, the two causes must be an action and an *omission* by an agent (as opposed to another action by the same agent). Second, the action and the omission must be *properly aligned,* that is, the omission in question must be the failure to act in the way that caused or would have caused E. This is to say, if one of the causes is the agent's φ-ing, then the other cause must be the agent's failing to φ (as opposed to, say, the agent's failure to ψ). Third, the action and the omission must be such that, in the scenarios where they obtain, they cause the same *token* outcome, not just outcomes of the same type.[16]

Is it possible to find a case that meets these three desiderata? In what follows, I look at two examples. The first example I will consider is Assassination. Once again, here is the case:

Assassination: Assassin shoots Victim and, as a result, Victim dies. However, Backup is waiting in reserve. Had Assassin not shot, Backup would have, and Victim would still have died (in a very similar way, at around the same time, etc.).

One might believe that this is a counterexample to CDM (A/O) because one might reason in the following way. Assassin's shooting caused Victim's death. However, had Assassin not shot, his failure to shoot would have caused Backup to shoot, and Backup's shooting would have in turn caused Victim's death. It follows by transitivity that Assassin's failure to shoot would *also* have caused Victim's death.

[16] These desiderata help distinguish CDM from clearly false theses in the vicinity. For instance, the claim that there cannot be more than one way to cause an outcome, and the claim that it is impossible for an action and the corresponding omission to cause outcomes of the same type.

In particular, one might think that Assassination meets the three desiderata. For, first, Backup would have acted as a result of one of Assassin's omissions. Second, the omission that Backup would have acted as a result of is precisely Assassin's failure to shoot, that is, the omission corresponding to the action that caused the death in the actual scenario. And third, Victim's death would have occurred in very much the same way if Assassin had shot or if he hadn't. Hence, the death if Assassin had shot and the death if he hadn't shot are presumably the same death.

In what follows I will argue that Assassination isn't a counterexample to CDM (A/O) because, while Assassin's shooting caused the death, his failing to shoot would *not* have caused the death (although it would have caused Backup to shoot, which would have caused the death). Hence, I will be arguing that we should reject the transitivity of causation at least in contexts of this type.

Let me pause here for a moment and remind you of the dialectic. This paper started out with an assumption: the assumption that the difference-making idea is worth pursuing. I said that I would be arguing that, *if* we wish to respect the difference-making idea, *then* we should endorse my view on how to cash it out (and I said that I would draw attention to an important advantage of embracing the difference-making idea at the end of the paper). I will now put this assumption to work in the following way. I will argue that the assumption that we should respect the difference-making idea is likely to lead us to say that the transitivity of causation fails in some contexts and, if it fails in those contexts, then it fails in Assassination. As a result, Assassination fails to be a counterexample to CDM (A/O).[17]

Let me start by reminding you of the following case:

> *Switch*: Victim is stuck on the railroad tracks. A runaway train is hurtling down the tracks when it approaches a switch. I flip the switch, and the train turns onto a side track. However, the tracks reconverge a bit further ahead, before the place where Victim is standing. Victim dies.

[17] The assumption that we should respect the difference-making idea plays an important role in my argument because it is not easy to argue against the transitivity of causation. Merely pointing to seeming counterexamples to transitivity does not seem to be enough, for the view that causation is transitive seems to be deeply entrenched in our way of thinking. When we look for the causes of a given event, we often proceed by tracing a causal chain back to earlier events and concluding that those earlier events are causes of the later event. This method assumes that causation is transitive. For discussion of transitivity, see Hall (2000) and (forthcoming), Hitchcock (2001), Paul (2000), and Yablo (2002) and (forthcoming).

In section 2, I pointed out that, to the extent that we wish to respect the difference-making idea, we should say that the flip isn't a cause of the death in Switch. For, intuitively, the flip didn't make a difference to the death.

Now, on the assumption that the flip isn't a cause in Switch, it seems that it isn't a cause in the following variant of Switch either:

> *Switch-with-Side-Track-Disconnected*: Again, I am by the switch but this time I see that part of the side track is disconnected. I think that I can make the train derail by turning it onto the side track. Hence, I flip the switch and the train turns. However, Backup is waiting by the side track. When he sees that I flip the switch, he rapidly reconnects the side track. The train runs on the side track for a while, then on the main track again, and finally kills Victim.

If anything, we feel even *more reluctant* to say that my flipping the switch is a cause of Victim's death in this case, where the side track was disconnected when I flipped the switch.

Notice that my claim about Switch-with-Side-Track-Disconnected is a *conditional* claim. What I am suggesting is that, *if* we said that the flip isn't a cause in Switch, *then* we would have to say that it isn't a cause in Switch-with-Side-Track-Disconnected. We might be prepared to say that the flip is a cause in Switch if, for instance, we held the view that determining the route to an event is sufficient for causing an event (which requires giving up the difference-making idea). If we held this view, then we would want to say that the flip is also a cause in Switch-with-Side-Track-Disconnected, since, in this case too, the flip determines the route to the death. My claim is only that, on the assumption that the flip isn't a cause in Switch, as the difference-making idea dictates, it is even more clearly not a cause in Switch-with-Side-Track-Disconnected.

Let me also stress that my claim about Switch-with-Side-Track-Disconnected is a purely *causal* claim and, as such, it is independent of any moral considerations. Just as we did with Switch in section 2, we can see that the intuitions about Switch-with-Side-Track-Disconnected are genuinely causal, and not merely moral, by imagining a similar scenario deprived of moral agents. Imagine, again, that what causes the switch to be flipped is a gust of wind, what reconnects the side track is a mechanism that is

automatically triggered when the switch is flipped, and what is lying on the tracks, and gets squashed by the train, is an apple. Still, we feel that, if the flipping of the switch isn't a cause of the outcome in Switch, where the side track was connected all along, then it is even more clearly not a cause in Switch-with-Side-Track-Disconnected, where the side track had to be reconnected in order for the train to reach the apple.

Now, the following also seems to be true about Switch-with-Side-Track-Disconnected: my flipping the switch caused Backup to reconnect the side track, and the reconnection of the track by Backup caused, in turn, Victim's death. It is intuitively clear that my flipping the switch caused the reconnection of the track by Backup, for the flip was the event that triggered that kind of behavior in Backup: Backup was determined to reconnect the side track just in case I flipped the switch, and he acted on that decision. And it is also intuitively clear that Backup's reconnecting the side track caused Victim's death, for the death would easily have been prevented otherwise: by reconnecting the track, Backup ensured that the death happened. Hence, my flipping the switch caused Backup to reconnect the side track, and Backup's reconnecting the side track caused Victim's death; however, on the standing assumptions about Switch and difference-making, my flipping the switch *didn't* cause Victim's death. This is to say, on the standing assumptions about Switch and difference-making, transitivity fails in Switch-with-Side-Track-Disconnected.[18]

I have argued that the assumption that we should respect the difference-making idea leads to the rejection of the transitivity of causation. For the difference-making idea dictates that my flipping the switch didn't cause Victim's death in Switch. Now, if my flipping the switch didn't cause Victim's death in Switch, then it probably didn't cause it in Switch-with-Side-Track-Disconnected either. But then it

[18] For some people, the failure of transitivity would arise earlier in my argument, in Switch itself. This would be so if we believed that, while my flipping the switch didn't cause the death, it caused the train to run on the side track, which caused the death. However, we needn't say this about Switch. Maybe what my flipping the switch caused wasn't what caused the death. Maybe what caused the death was the event of the train's running towards Victim, and what my flipping the switch caused was the fact that such an event had a certain feature, or the fact that it happened in a certain way, i.e., the fact that it happened on the *side* track. L. Paul rebuts some alleged counterexamples to transitivity in this way in Paul (2000) (see also Thomson (2003)). Notice, however, that Switch-with-Side-Track-Disconnected doesn't seem to be open to the same treatment: it seems clear that my flipping the switch caused Backup to reconnect the track, and that this caused the death.

seems that we should say that, in Switch-with-Side-Track-Disconnected, my flipping the switch caused Backup to reconnect the side track, which caused the death, but my flipping the switch didn't cause the death. In what follows, I argue that the scenario where Assassin fails to shoot in Assassination is on a par with the scenario where I flip the switch in Switch-with-Side-Track-Disconnected. Hence, if transitivity fails in Switch-with-Side-Track-Disconnected when I flip the switch, it also fails in Assassination when Assassin fails to shoot.

I have pointed out that my flipping the switch is even more clearly *not* a cause of Victim's death in Switch-with-Side-Track-Disconnected than in the original case, Switch. Why is this? Intuitively, this is because my flipping the switch only made it *more difficult* for the death to happen, by calling for Backup's intervention. Given that I flipped the switch, Backup had to intervene or else the death wouldn't have happened, while, had I not flipped the switch, the death would have occurred much more easily, without the need for Backup's intervention.[19] Now, I will argue that the relation between Assassin's failure to shoot and Victim's death in Assassination is significantly similar to the relation between my flipping the switch and Victim's death in Switch-with-Side-Track-Disconnected. Hence, if my flipping the switch did not cause the death in Switch-with-Side-Track-Disconnected, Assassin's failure to shoot wouldn't have caused the death in Assassination.

To see this, imagine that Assassin failed to shoot in Assassination. So Backup shot, and Victim died. Then, just as in Switch-with-Side-Track-Disconnected, Assassin's failure to shoot only made it more difficult for Victim's death to happen, by calling for Backup's intervention. Given that Assassin didn't shoot, Backup had to intervene or else the death wouldn't have happened, while, had Assassin shot, the death would have occurred much more easily, without the need for Backup's intervention. This suggests that the same reasons that should lead us to reject transitivity in Switch-with-Side-Track-

[19] Here is another example in the same vein. While swimming in the sea, a child is attacked by a shark. The child is then rushed to a hospital, where he is treated for a few days, until his wounds heal. Intuitively, the shark attack caused the medical treatment, the medical treatment caused the child's good health, but the shark attack did *not* cause the good health. This is so because, intuitively, the shark attack only made it more difficult for the child's good health to ensue, given that it introduced the need for the medical treatment.

Disconnected (in the scenario where I flip the switch) should also lead us to reject transitivity in Assassination (in the scenario where Assassin fails to shoot). They should lead us to say that, while Assassin's failure to shoot would have caused Backup to shoot, and while Backup's shooting would have caused Victim's death, Assassin's failure to shoot would *not* have caused Victim's death. If so, Assassination fails to be a counterexample to CDM (A/O) because it is not true that *both* Assassin's shooting *and* Assassin's failing to shoot would have caused Victim's death.

My diagnosis of Assassination can be generalized to cases of the following sort. An agent's action and the corresponding omission would both have been followed by a certain outcome E. The agent's action is the sort of action that normally leads to outcomes of E's type, and in the actual case it leads to the outcome in the normal way. The agent's omission, by contrast, is the sort of omission that could only lead to the outcome via an abnormal route, which contains the intervention of a backup mechanism without which the outcome wouldn't have occurred. As a result, it seems wrong to count the omission as a cause of the outcome in the scenario where the omission obtains. Hence, it is not the case that both the action and the omission would have caused the same outcome. Hence, cases of this type aren't counterexamples to CDM (A/O).[20]

In cases of the type that we have just examined, one of the candidates for being a cause, the agent's action, has an initial causal advantage over the other candidate, the agent's omission, and thus, it is a better *prima facie* candidate for being a cause. But, what about cases where the two candidates are intuitively *on a par?* That is, what about cases where neither candidate is a better *prima facie* candidate for being a cause? Couldn't cases of this type be counterexamples to CDM (A/O)? I turn to a case of this type in the next section.

---

[20] In principle, the same style of reasoning should apply to the flipside of this case: a case where the route containing the action is *less* straightforward and the route containing the omission is *more* straightforward. However, as we will see in chapter 2, I think that there is an asymmetry between actions and omissions in this respect. I think that, if the outcome would have occurred anyway in the absence of the omission, the omission isn't a cause, *regardless* of how straightforward the actual route to the outcome is. I discuss the connections between the results of this and the next chapter at the end of the next chapter.

5. Argument for CDM (Part II)

Here is such a case:

> *Two-Assassins*: I hired two assassins and I gave them the following instructions. Assassin 1 is to shoot Victim just in case I nod at *t*. Assassin 2 is to shoot Victim just in case I fail to nod at *t*. As a matter of fact, I nod at *t*, Assassin 1 shoots and Victim dies.

Someone might want to say that this is a counterexample to CDM (A/O) for reasons parallel to those mentioned in our discussion of Assassination. Namely, my nodding caused Assassin 1 to shoot, which caused the death; hence, it is tempting to say that my nodding caused the death. Similarly, my failure to nod would have caused Assassin 2 to shoot, which would have caused the death; hence, it is tempting to say that my failure to nod would also have caused the death.

One might also think that Two-Assassins meets the desiderata from the last section. First, the two candidate causes are an agent's action and an agent's omission. Second, the action and the omission are properly aligned: Assassin 2 would have shot just in case I failed to nod, where my nodding is precisely that which made Assassin 1 shoot in the actual scenario. Third, we can fill in the details of the case so that the death that Victim would have encountered if Assassin 2 had shot would have been the same death as the one that he encountered given that Assassin 1 shot (we can assume that the deaths would have occurred at around the same time, and in a very similar fashion).

Finally, Two-Assassins is a case where the agent's action and the agent's omission are intuitively on a par with respect to their causal powers: it seems that one of them is a cause just in case the other is a cause. Hence, my argument against Assassination from the last section doesn't apply to Two-Assassins.

I will argue that Two-Assassins fails to be a counterexample to CDM (A/O) because (on the standing assumptions about Switch and difference-making) transitivity fails in this case as well. However, my diagnosis of Two-Assassins will differ from that of Assassination, in the following way. I have claimed that, in Assassination, while Assassin's shooting caused the death, his failure to shoot wouldn't have caused it. By contrast, I will claim that, in Two-Assassins, *neither* my nodding *nor* my

failure to nod would have caused the death. This is to say, I will argue that, in a case where the agent's action and the omission are intuitively on a par, neither is a cause of the outcome.

Again, my argument will be based on an analogy with a variant of Switch. The variant that we need now is one where, not just one, but the two tracks are initially disconnected:

> *Switch-with-Both-Tracks-Disconnected*: This time, both of the tracks are disconnected after the switch. However, there is one assassin next to each track. If I don't flip the switch, Assassin 1 will reconnect the main track and Victim will die. If I flip the switch, Assassin 2 will reconnect the side track and Victim will die.

Suppose that I flip the switch. Consequently, the train turns onto the side track, which Assassin 2 rapidly reconnects, then the tracks reconverge, and Victim dies. Did my flip cause the death? If it didn't cause it in the original case, Switch, it seems that it didn't cause it in this case either. Intuitively, in neither case did the flip make a difference to the death, for the actual scenario and the scenario where I don't flip the switch are relevantly parallel: whereas Assassin 2 reconnects the track in the actual case, Assassin 1 reconnects the track in the case where I don't flip the switch. Hence, it seems that, if the flip didn't cause the death in Switch, then it didn't cause the death in Switch-with-Both-Tracks-Disconnected either. But it is clear that, in Switch-with-Both-Tracks-Disconnected, my flipping the switch caused Assassin 2 to reconnect the side track, which in turn caused the death. Hence, on the standing assumptions about Switch and difference-making, transitivity fails in Switch-with-Both-Tracks-Disconnected if I flip the switch.

Alternatively, suppose that I *don't* flip the switch. The train then continues along the main track, which Assassin 1 rapidly reconnects, and Victim dies. Again, it seems that, if my failure to flip the switch didn't cause the death in Switch, then it didn't cause it here either, even though it caused Assassin 1 to reconnect the track, which in turn caused the death. Thus, on the standing assumptions about Switch and difference-making, transitivity fails *both* if I flip the switch *and* if I don't flip the switch.

Now, Two-Assassins strikes me as on a par with Switch-with-Both-Tracks-Disconnected. Just as neither my flipping the switch nor my failing to flip the switch would have made a difference to Victim's

death in Switch-with-Both-Tracks-Disconnected, it seems that neither my nodding nor my failing to nod would have made a difference to Victim's death in Two-Assassins. For, in Two-Assassins too, the scenario where I nod and the scenario where I fail to nod are relevantly parallel: whereas Assassin 1 shoots in the case where I nod, Assassin 2 shoots in the case where I don't nod. Thus, if neither the flip nor the failure to flip would have caused the death in Switch-with-Both-Tracks-Disconnected, then, similarly, neither my nodding nor my failure to nod would have caused the death in Two-Assassins. My nodding wouldn't have caused the death, even though it would have caused Assassin 1 to shoot, which would have caused the death. And my failure to nod wouldn't have caused the death, even though it would have caused Assassin 2 to shoot, which would have caused the death.

Now, one might find this puzzling. *I* hired Assassin 1 and gave him the instruction to shoot just in case I nodded. How can I say, then, that my nodding wouldn't have caused the death? Similarly, *I* hired Assassin 2 and gave him the instruction to shoot just in case I didn't nod. How can I say, then, that my failure to nod wouldn't have caused the death?

To see that this isn't a problem, imagine that I also hired the two assassins standing by the tracks in Switch-with-Both-Tracks-Disconnected. Thus, imagine that I gave Assassin 1 the instruction to reconnect the main track in case I didn't flip the switch, and Assassin 2 the instruction to reconnect the side track in case I flipped the switch. This doesn't change the verdict about the causal powers of my flipping the switch, or of my failure to flip the switch. To be sure, if I hired the two assassins, I caused the death. But I caused it in virtue of my *hiring the two assassins,* not in virtue of flipping the switch or failing to flip it.[21] By hiring the two assassins and giving them the instructions that I gave them, I made sure that the death would happen. But I also made sure that, at the time when I had to decide whether to

[21] How did my hiring the two assassins cause the death? By starting a causal route to the death, which included the reconnection of the side track by the assassin on that track. Rejecting transitivity is consistent with saying this. More generally, rejecting transitivity is consistent with saying that, in order for C to be a non-immediate cause of E, there must be an intermediary, D, that is caused by C and that causes E.

flip the switch or not, what I decided to do then couldn't make a difference. This is to say, I made sure that nothing I did or failed to do at that moment could count as a cause of the death.[22]

Similarly, my claim is that, in Two-Assassins, I caused Victim's death but not in virtue of nodding, or failing to nod. I caused the death by hiring the two assassins and giving them the specific instructions that I gave them. By hiring the assassins and giving them those instructions, I made sure that the death would happen, but I also made sure that, at the time when I had to decide whether to nod or not, what I did then couldn't make a difference. This is to say, I made sure that nothing I did or failed to do at that moment could count as a cause of the death.

Let me sum up the results of the last two sections. I have argued that, on the assumption that we should respect the difference-making idea, it follows that two main attempts at counterexamples to CDM (A/O), Assassination and Two-Assassins, fail. I first argued that Assassination isn't a counterexample to CDM (A/O), for, while the agent's action is a cause of the outcome, the agent's omission wouldn't have been a cause of the outcome. Then I pointed out that Assassination is a case where, intuitively, one of the candidate causes, the agent's action, has an initial advantage over the other, the agent's omission. So the natural reaction was to look instead for a case where the action and the omission are intuitively on a par with respect to their causal powers and to see whether a case of that sort has a better chance of being a counterexample to CDM (A/O). This is how we arrived at Two-Assassins. I argued, however, that, as a result of making the action and the omission equally good candidate causes, as in Two-Assassins, it turns out that *neither* is a cause, not that both are. This is to say, by depriving the candidate causes of any initial advantage over each other, we deprive them of causal power altogether. Hence, Two-Assassins also fails to be a counterexample to CDM (A/O).

My argument for CDM (A/O) takes, then, the following form. Possible counterexamples to CDM (A/O) can be grouped into two main classes: the class of cases where the action and the omission aren't

[22] What if I could call the whole thing off by, say, waving my hand in a particular way? Then my failure to wave my hand in that way would have been a cause of Victim's death. Still, my flipping the switch or my failing to flip the switch wouldn't have been a cause of the death.

intuitively on a par (with respect to their causal powers) and the class of cases where the action and the omission are intuitively on a par. My discussion of Assassination suggests that the cases in the *first* class fail because *only one* of the candidate causes is a genuine cause. In turn, my discussion of Two-Assassins suggests that the cases in the *second* class fail because *neither* of the candidate causes is a genuine cause. Since any alleged counterexample will fall into one of the two classes, and since the cases I have discussed seem to be representative of their class, I conclude that there is good reason to believe that CDM (A/O) is true. More precisely, I conclude that there is good reason to believe that CDM (A/O) succeeds in capturing the difference-making idea in the case of actions and omissions of agents.

Finally, my focus on actions and omissions of agents was only for simplicity. In principle, it should be possible to use the same style of argument to show that the general claim, CDM, is true. I conclude that CDM succeeds in cashing out the difference-making idea: if causes are difference-makers, it is in virtue of the fact that events and their absences would not have caused the same effects.

6. Implications for moral responsibility

In this final section, I draw attention to some important results that CDM has for issues in moral responsibility.

I will suggest that CDM achieves a particularly nice fit between the concepts of causation and moral responsibility. One way in which this emerges is as follows. Ordinarily, we regard ourselves as morally responsible for the (foreseeable) consequences of what we do or fail to do. Intuitively, this seems to be because, ordinarily, we regard ourselves as having a *choice* whether to cause those consequences. I say "ordinarily," because there are some *extraordinary* circumstances where this is not the case. Notably, if we are coerced to behave in certain ways, or if we act under the influence of some powerful drug, then we might not have a choice whether to cause the ensuing consequences and thus we might not be responsible for them. But these cases are extraordinary in that they are cases where we lack control of the *actions and omissions* that issue in those consequences and, correspondingly, they are cases where we do not have a choice whether to cause those consequences.

Now, according to CDM, and, in particular, according to CDM (A/O), *whenever* we have a choice whether to act or fail to act in certain ways, we *thereby* have a choice whether to cause the ensuing consequences. According to CDM (A/O), it simply couldn't be that, both by acting and by failing to act, I would be causing the same consequences. Hence, CDM (A/O) suitably fits the way in which we ordinarily think of ourselves as responsible for the consequences of our actions and omissions.

By contrast, imagine what we would have to say if we rejected CDM (A/O). If we rejected CDM (A/O), then we would have to say that, on some occasions, both acting in certain ways and failing to act in those ways would cause the same outcomes. Then, on those occasions, we wouldn't have a choice whether to cause those outcomes. For, in those cases, regardless of what we did (were we to act in certain ways or were we to fail to act in those ways), we would be causing those outcomes. Moreover, in those cases we wouldn't have a choice whether to cause certain outcomes *even if we happened to be in complete control of the actions and omissions that caused the outcomes.* But, as I have pointed out, barring exceptions in which we lack control of our own actions and omissions, we tend to regard ourselves as responsible for the consequences of our actions and omissions because we tend to regard ourselves as having a choice whether to cause those consequences. Hence, rejecting CDM (A/O) would clash with the ordinary way in which we think of ourselves as responsible for the consequences of our actions and omissions.

Let me illustrate this point with an example, before moving on to my last remark. Take Assassination. Suppose that I have the choice between shooting and failing to shoot. Intuitively, I then have a choice whether to cause Victim's death. Regardless of whether Backup shoots, and regardless of whether Backup shoots as a result of my failing to shoot or because he was going to shoot anyway, the intuitive thought is that, if I don't shoot, I don't cause the death, and therefore I am not responsible for the death. Now, someone might be prepared to give up this thought, upon realizing that it requires rejecting the transitivity of causation in some contexts. What I am suggesting is that this would come at a cost: the cost of giving up the ordinary way in which we regard ourselves as responsible for the consequences of our actions and omissions. CDM takes the opposite route: it allows for the intransitivity of causation in

some contexts, but it accommodates the ordinary way in which we regard ourselves as responsible for the consequences of our actions and omissions.

Finally, I will suggest that CDM is particularly helpful in accounting for the lack of moral responsibility of agents in some cases of moral luck. Briefly, a case of moral (good) luck is a case where an agent that behaves in a morally wrong way doesn't come out responsible for a harm thanks to the obtaining of some circumstances that are outside of the agent's control.[23] Here is a case of moral luck with respect to which CDM can prove particularly useful:

> *Switch-with-Main-Track-Unexpectedly-Connected*: Again, Victim is trapped on the tracks. I want Victim to die, and I have reason to believe that the main track is disconnected. So, thinking that the train will derail if it continues on the main track, I flip the switch. As it turns out, however, the main track has never been disconnected. As a result of my flipping the switch, the train turns onto the side track, but then the tracks reconverge and the train hits Victim.

Intuitively, this is a case of moral luck because, even if I acted wrongly in flipping the switch, I am not responsible for Victim's death (I might be responsible for intending to cause his death, for trying to cause his death, etc., but not for the death itself). I *thought* that I would cause Victim's death by flipping the switch, and I *intended* to cause the death by flipping the switch. However, even if the death did occur, it seems that, given that the main track was connected all along, my flipping the switch did not cause the death. Switch-with-Main-Track-Unexpectedly-Connected only differs from the original case, Switch, in what I *thought* was the case, not in what was actually the case. Hence, if my flipping the switch does not cause the death in Switch, it does not cause it in Switch-with-Main-Track-Unexpectedly-Connected either.

Now, as I have suggested, CDM (together with the observation that the contribution that the flip made to the death is on a par with the contribution that the failure to flip would have made to the death) entails that the flip did not cause the death in Switch. For the same reason, CDM entails that the flip did

---

[23] See Nagel (1979), ch. 3, and Williams (1981), ch. 2. All of Nagel's and Williams' examples are cases where the agent isn't responsible for a harm that doesn't occur but could easily have occurred. By contrast, I will focus on cases where the harm does occur but the agent still doesn't cause it.

not cause the death in Switch-with-Main-Track-Unexpectedly-Connected. As a result, CDM helps us explain my moral luck in this case.[24]

It is worth noting that, on many theories of causation, it would turn out that, on the contrary, my flipping the switch *did* cause the death in Switch-with-Main-Track-Unexpectedly-Reconnected. By way of example, note that my flipping the switch helped to determine the causal route to the death, by determining which track the train would take. Hence, in particular, theories of causation according to which helping to determine the causal route to an outcome is sufficient for causing the outcome would entail that my flipping the switch *is* a cause of the death. As a result, these theories would fail to account for my moral luck in cases of this type.[25]

I conclude that, not only does CDM succeed in capturing the difference-making idea, but it also yields a concept of cause that has the seemingly right kinds of connections to moral concepts, such as the concept of moral responsibility.

7. Conclusions

In this paper I argued for a particular way of cashing out the idea that causes are difference-makers. I argued that we should interpret this idea in the following way: a cause makes a difference to its effect in that, if it hadn't occurred, the absence of the cause *wouldn't* have been a cause of the effect. I also argued that this view of causation has important implications for moral responsibility, in particular, I argued that it carves up a concept of cause that is particularly well suited for the work that causation does in moral theory.

---

[24] What principle warrants the inference "The flip wasn't a cause of the death, hence I am not morally responsible for the death"? I answer this question in chapter 3. As we will see then, I *don't* think it's the principle that responsibility requires causation, for I believe that this principle is false.

[25] Typically, intentionally causing a harm (in a "non-deviant" way) is taken to be sufficient for being morally responsible for the harm. (See, e.g., Feinberg (1970).) If this is so, any theory of causation that entails that I caused the death will have serious trouble explaining my moral luck. For, if I caused the harm by flipping the switch, I did it intentionally (and in a "non-deviant" way, i.e., by making the train turn onto the side track, as I intended). As a result, if we said that my flipping the switch caused the death, it is likely that we would have to revise our views on moral responsibility.

Chapter 2

A New Asymmetry Between Actions And Omissions

1. Introduction

Consider the following case:

> *Shooting*: I freely decide to shoot Victim. I pull the trigger and Victim dies. Had I wavered in my decision, an evil neuroscientist who has been secretly monitoring my brain and who can, at any time, take control of my decision-making processes, would have sent a signal to my brain that would have made me decide to pull the trigger anyway.

Cases of this type are famous in the literature on moral responsibility; they are called "Frankfurt-style" cases.[26] Now, compare Shooting to this other case:

> *Sharks*: While walking by the beach, I see a child drowning. I think I could jump into the water and save him but I deliberately refrain from doing so. The child drowns. Unbeknownst to me, the water is infested by sharks. Had I jumped in, the sharks would have attacked me and prevented me from saving the child.[27]

There is an interesting difference between Shooting and Sharks. In Shooting, I freely made the decision to shoot and I freely acted on that decision. Even though the neuroscientist would have prevented me from deciding to do anything else if I had even tried to decide differently, as a matter of fact he didn't have to intervene. Thus, I am responsible for Victim's death. By contrast, in Sharks, I am not responsible for the child's death. Again, I freely decided to stay on the shore, and the sharks didn't have to intervene, but, somehow, the fact that I couldn't have saved the child given that the water was infested by sharks seems to relieve me of responsibility for the death.[28]

---

[26] After H. Frankfurt (Frankfurt (1969)). Frankfurt uses examples of this type to attack the claim that moral responsibility requires the ability to do otherwise.

[27] A case by J. M. Fischer and M. Ravizza (Fischer and Ravizza (1998), p. 125).

[28] Throughout the paper, I will reserve the word "responsible" for *morally* responsible. Also, I will restrict my attention to responsibility of agents for bad outcomes or for actions and omissions resulting in bad outcomes.

Note that the claim is only that, in Sharks, I am not responsible for the *death* of the child. Presumably, I am still a bad *person,* given that I thought that I could save the child but, even so, I didn't jump in. (In fact, I am probably as bad a person as someone who decides not to step in when he *could* have saved the child.)[29] Also, in Sharks, I *am* responsible for *something*—presumably, for deciding to stay on the shore, and for not trying to prevent the child's death. But I am not responsible for the child's *death.* (If you are not convinced, imagine that, in addition to the sharks, there is an impenetrable wall of rocks in the water, and a big net, and all kinds of obstacles, each one of which would have prevented me from saving the child. It would be very implausible to say that I am responsible for the child's death in those circumstances.)[30]

In light of the contrast between cases like Shooting and Sharks, some philosophers have suggested that there is a moral asymmetry between actions and omissions. For we can rephrase the difference between Shooting and Sharks in terms of actions and omissions, as follows. In Shooting, I am responsible for my action of *killing Victim,* even though, given the presence of the neuroscientist, I couldn't have failed to kill Victim. By contrast, in Sharks, I am not responsible for my omission of *failing to save the child,* for, given the presence of the sharks, I couldn't have saved the child. As a result, it has been suggested that the following thesis holds—henceforth, the *Old Asymmetry* thesis:

> OA: Whereas an agent can be responsible for an *action* even if he couldn't have done otherwise, an agent cannot be responsible for an *omission* if he couldn't have done otherwise.[31]

---

[29] Here I am relying on a common distinction between judgments about character and judgments about responsibility for outcomes.

[30] Someone might point out that I still *chose* not to save the child in those circumstances. But, can I choose to omit to do something that I couldn't have done? Imagine that I am superstitious and think that I can save my moribund enemy by performing a healing ritual. I refrain and my enemy dies. Did I choose to fail to save him? It seems not. In any case, it is clear that I am not responsible for his death.

[31] See, e.g., van Inwagen (1978).

In other words, OA claims that, whereas one can be responsible for *acting* a certain way even though one couldn't have failed to act that way, one cannot be responsible for *failing to act* a certain way if one couldn't have acted that way.

In this paper I discuss the question whether there is a moral asymmetry between actions and omissions. I argue that there *is* a moral asymmetry, but it is *not* OA. I offer a new asymmetry thesis, and I argue that this new asymmetry thesis succeeds in capturing the moral asymmetry between actions and omissions that is illustrated by the contrast between Shooting and Sharks.

The plan for the paper is the following. The first part is about OA. As I said, I will reject OA; however, it is important to reject it for the *right* reasons. As we will see, there are people who have rejected OA for the wrong reasons. Those people are confused, not in their belief that OA is false, but in the reasons they have for holding that belief. Thus, in the next section, I explain what the *wrong* reasons for rejecting OA are, and why they are wrong. The discussion in that section centers around the conditions of transmission of responsibility. In section 3, I look into this issue in more detail. I put forth a principle of transmission of responsibility that has causation as an essential ingredient and that serves as a springboard for the next two sections. In section 4, I explain how this principle supports the *right* reasons for rejecting OA, and, in section 5, I explain how it gives rise to a new moral asymmetry between actions and omissions. The new moral asymmetry rests on a *causal* principle according to which actions and omissions have different causal powers. I argue for this principle in sections 6 and 7. As a result, the picture that emerges from the paper is the following: there is a moral asymmetry between actions and omissions, and this moral asymmetry is the consequence of a causal asymmetry between actions and omissions and the role causation plays in the transmission of responsibility.

A note of clarification. Throughout the paper, I will be assuming that omissions can cause things just as actions can. Thus, my proposal will *not* be that there is a moral asymmetry between actions and omissions because actions can be causes while omissions can't.[32] I will argue that, although omissions

[32] In particular, my proposal will not be that omissions can never make us responsible for anything because they can never cause anything (for an example of this view, see Weinryb (1980)). One reason to allow for the existence of

can be causes, the conditions under which they are causes are different from the conditions under which actions are causes, and this results in a moral asymmetry between actions and omissions.

2. The *wrong* reasons for rejecting OA

Recall OA:

> OA: Whereas an agent can be responsible for an *action* even if he couldn't have done otherwise, an agent cannot be responsible for an *omission* if he couldn't have done otherwise.

Some people have rejected OA for the wrong reasons. They have reasoned as follows.[33] As we have seen, there are Frankfurt-style action cases in which an agent is responsible for his action despite the fact that he couldn't have done otherwise. Shooting was a case of that type: in Shooting, I am responsible for killing Victim although, given the presence of the neuroscientist, I couldn't have done otherwise. Similarly, the reasoning goes, there are Frankfurt-style *omission* cases in which an agent is responsible for his omission even though he couldn't have done otherwise. This is an alleged example:

> *Frankfurt-style omission case* (FSOC): I see the child drowning, I think I can save him by jumping into the water, but I freely decide not to jump in. This time there are no sharks in the water, but the evil neuroscientist is monitoring my brain. Had I wavered in my decision, he would have made me decide not to jump in.

Those who have rejected OA for the wrong reasons have thought that, in FSOC, I am responsible for my failure to save the child, even though (given the presence of the neuroscientist) I couldn't have saved him. Thus, they have concluded that it is possible for an agent to responsible for an omission when he couldn't have done otherwise, and so OA is false.

---

causation by omission is that, arguably, if we didn't allow for it, it would follow that there is also a lot less "positive" causation than we normally think there is (for an argument for this view, see Schaffer (2000)).

[33] See, e.g., Clarke (1994) and McIntyre (1994).

I will dub this objection to OA, the *Frankfurt-style objection* to OA, and I will dub someone who raises this kind of objection to OA, a *Frankfurt-style objector* to OA. In what follows, I argue that the Frankfurt-style objection to OA is flawed.

The first thing to notice is that the Frankfurt-style objector to OA cannot just appeal to intuition to support his claim that I am responsible for my failure to save the child in FSOC. For the question that arises naturally is: how can I be responsible for my failure to save the child in FSOC (as the Frankfurt-style objector to OA claims), if (as everybody seems to agree) I am not responsible in Sharks? In both cases, I couldn't have saved the child. Then, why am I, according to the Frankfurt-style objector, responsible for my failure to save the child in FSOC but not in Sharks?

The Frankfurt-style objector will probably try to say the following.[34] In Sharks, had I decided to jump in, I still wouldn't have saved the child (because the sharks would have stopped me). In FSOC, by contrast, had I decided to jump in, I *would* have saved the child (because there aren't sharks or any other obstacles in the water). True, I couldn't easily have decided to jump in, since the neuroscientist was determined not to let me make that decision. Still, *had* I decided to jump in, I would have saved the child. That is, in the closest possible world where I decide to jump in—a world where the neuroscientist fails and where there are no obstacles in the water—I save the child. In other words, in FSOC, whether the child lived or died *hinged on* what I decided to do. And in all of these cases I am responsible for making the decisions that I made (no one forced me to make them). Presumably, then, I am also responsible for my failure to save the child in FSOC. Or so the Frankfurt-style objection to OA goes.

I take it that the thought is that, given that the child's death depended on my decision, for which I am responsible, I am responsible for the child's death. And thus, given that I am responsible for the child's death, I am responsible for my failure to save the child. This seems like a natural thing for the Frankfurt-style objector to say. So I suggest the following reconstruction of the argument by the Frankfurt-style objector:

---

[34] Both Clarke and McIntyre suggest an argument along these lines in Clarke (1994) and McIntyre (1994).

(1) In FSOC, I am responsible for my decision not to jump in.

(2) The child's death depended on that decision.

(3) If I am responsible for X, and an outcome Y depends on X, then I am responsible for Y.

(4) Therefore, I am responsible for the child's death.

(5) If I am responsible for the child's death, then I am responsible for my failure to save the child.

(6) Therefore, in FSOC, I am responsible for my failure to save the child.

In what follows, I am not going to discuss step (5). I think it is intuitively plausible that, in a case where I fail to save the child, if I am responsible for the child's death, then I am also responsible for my failure to save him.[35] So I will assume that all the Frankfurt-style objector has to prove to make his case is that I am responsible for the child's death in FSOC.

A note about (3) is in order. As it stands, (3) is not very plausible, for it makes agents come out responsible for things that they couldn't possibly have foreseen. To make it more plausible, we should probably add more provisos; in particular, we should probably say: "If I am responsible for X, an outcome Y depends on X, *and it was foreseeable that Y (or an outcome of Y's type) was likely to follow X,* then I am responsible for Y." I will not be concerned with any extra provisos because it is likely that they will be met in FSOC (and in other cases I will focus on) and thus, they will not matter for my purposes here. For instance, in FSOC, it *was* foreseeable that the child's death would likely follow my decision not to jump in. I will assume that (3) can be suitably revised in this respect, and I will leave the revisions implicit in what follows.

I will call the argument above an *argument from dependence.* That argument relies heavily on a certain concept of dependence that allegedly links my *decision not to jump in* to *the child's death* in

[35] This is also the natural way to look at cases in which I am *not* responsible. For instance, as I pointed out in section 1, it is natural to think that, in Sharks, I am not responsible for my failure to save the child because I am not responsible for the child's death.

FSOC, and that purportedly transmits my responsibility from one onto the other. I will argue, however, that there is no concept of dependence that can do that and thus, the argument fails. The discussion that follows will center on the issue of the transmission of responsibility. This topic will play a fundamental role in the formulation of the new asymmetry later, so it is important, not only because of its implications for the Frankfurt-style objection to OA, but also because of its implications for the new asymmetry.

What can be the concept of *dependence* that the Frankfurt-style objector has in mind and that purportedly serves to transmit responsibility to the ensuing outcome? A natural candidate is *causation.* For it is natural to regard causation as a link that agents have with the world, in virtue of which they can be responsible for what happens in it. In other words, it is natural to regard causation as the means by which the responsibility of agents for their actions and omissions *transmits* to outcomes in the external world. In the next section I will argue that, not only is this a natural picture of how the responsibility of agents is transmitted to outcomes, but it is also the *right* picture. Now, however, I will argue that the Frankfurt-style objector cannot appeal to the concept of causation in his argument from dependence against OA.

If dependence is causation, then the relevant premises of the objector's argument about FSOC read:

(2') My decision not to jump in caused the child's death.

(3') If I am responsible for X, and X caused an outcome Y, then I am responsible for Y (with the usual provisos).

But (2') is false. Arguably, what caused the child's death is not my *decision not* to jump in, but my *failure to decide* to jump in. For the child didn't die in FSOC in virtue of what I *did* decide to do; he died in virtue of what I *didn't* decide to do.[36]

[36] Recall that I am assuming that omissions can be causes. But, what if the Frankfurt-style objector held a much *narrower* conception of causation according to which omissions cannot be causes? Or, what if he held a much *more permissive* conception of causation according to which my failure to decide to jump in is a cause, but my decision not to jump is *also* a cause? For reasons that will become apparent in the next section, if he said any of these two

Let me explain. First of all, it is important to distinguish between my decision not to jump in and my failure to decide to jump in. Arguably, these things are not identical. For they have different obtaining conditions: my failure to decide to jump in could have obtained without my decision not to jump in obtaining. For instance, I would have failed to decide to jump in if I hadn't made any decision at all but had kept deliberating what to do until the child drowned.[37] The question arises, then: which one of the two, my decision not to jump in or my failure to decide to jump in, has a better claim to cause the child's death in FSOC? Presumably, the latter. For one, the child would still have died if my failure to decide to jump in had obtained without my decision not to jump in obtaining (if I had remained undecided). Also, the fact that I decided not to jump in seems relevant to the child's death *only to the extent that* it entails that I didn't decide to jump in (if I decided not to jump in, I couldn't have decided to jump in, and thus, I couldn't have saved the child). This suggests that what is causally relevant to the child's death is the fact that I didn't decide to jump in, not the fact that I decided not to jump in.

Here is another example to illustrate this. Imagine that you have been poisoned and that you need to be injected with an antidote in the next second or you will die. Although I could give you the antidote, I give you a placebo. As a result, you die. However, you don't die as a result of my giving you the *placebo*; rather, you die as a result of my *not* giving you the *antidote*. The only respect in which my giving you the placebo seems relevant to your death is that, in the circumstances, if I gave you the placebo, I couldn't have given you the antidote. Thus, it seems that it is my not giving you the antidote, not my giving you the placebo, that caused your death. Similarly, in FSOC, it is my not deciding to jump in, not my deciding not to jump in, that caused the child's death.

---

things, then it would follow that, contrary to what we are assuming, causation *isn't* appropriate to transmit responsibility, and thus a different premise of the argument, (3'), would fail.

[37] Notice that I am rejecting the view that some philosophers have taken, according to which omissions are a subclass of actions. I do not find this view compelling. One serious problem with it is that sometimes, given an omission, there aren't *any* actions that can plausibly be identified with it, or, alternatively, there isn't a *single* action that can plausibly be identified with it. For discussion, see Weinryb (1980).

I have argued that premise (2') of the Frankfurt-style argument against OA fails: the child's death wasn't caused by my decision not to jump in, but by my failure to decide to jump in. Now, in light of this objection, the Frankfurt-style objector might think of substituting "my failure to decide to save the child" for "my decision not to save the child" in the relevant premises of the argument, thus:

(1') In FSOC, I am responsible for my failure to decide to jump in.

(2'') My failure to decide to jump in caused the child's death.

As I have argued, (2'') is probably true. But now the problem is that (1') is, at the very least, controversial. What *is* clear about FSOC is that I am responsible for what I *decided* (a mental *action* of mine); it is not equally clear that I am responsible for what I *failed* to decide (a mental *omission* of mine).

Let me explain. Premise (1) of the original Frankfurt-style argument was uncontroversial because my decision not to jump in is an *action* and because, as Frankfurt-style *action* cases show, I can be responsible for actions that I couldn't have avoided when I make them freely (when no one forces me to make them). But the revised premise, (1'), is *not* supported by a similar reasoning. True, no one forced me to fail to decide to jump in either. But this isn't enough to show that I am responsible for failing to decide to jump in. For, as we have seen, omission cases behave differently from action cases. In Sharks, for instance, no one forced me to fail to save the child, but I am still not responsible for failing to save the child.[38] If the Frankfurt-style objector to OA wants to insist that, in FSOC, I am responsible for failing to decide to jump in (not just for deciding not to jump in), then his argument becomes question-begging. For there is no more reason to believe that I am responsible for failing to decide to jump in (his premise) than there is to believe that I am responsible for failing to save the child (his conclusion), or, at least, the Frankfurt-style objector hasn't given us any such reason. After all, failing to decide to jump in and failing to save the child are both omissions, and, given the presence of the neuroscientist in the background, I could have avoided neither of them.

[38] I made a similar point in n. 30 above.

In this section I argued that the Frankfurt-style objection to OA fails, on the assumption that causation is the type of dependence that transmits the responsibility of agents for their actions and omissions to outcomes in the world. In the next section, I will argue that this assumption is very likely to be true. This will close my discussion of the Frankfurt-style objection to OA, and it will also get us a step closer to the new asymmetry.

3. The role of causation in the transmission of responsibility

If not causation, then what other type of dependence could transmit responsibility to outcomes? Maybe *counterfactual dependence* could? In what follows, I argue that counterfactual dependence isn't a suitable candidate. Counterfactual dependence will be my specific target, but, as we will see, my argument is likely to extend to other possible candidates as well. I will conclude that *causation* is the most suitable candidate for transmitting responsibility to outcomes.

Counterfactual dependence is defined in the following way:

> Y counterfactually depends on X just in case, had X not occurred, Y would not have occurred.

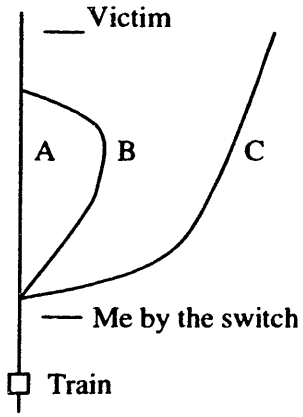
The principle according to which counterfactual dependence transmits responsibility to outcomes is the following:

> TR (Counterfactual): If I am responsible for X, and an outcome Y counterfactually depends on X, then I am responsible for Y (with the usual provisos).

Notice that, if TR (Counterfactual) were true, it could help the Frankfurt-style objector in his argument against OA. For it is possible to fill in the details of FSOC so that the child's death counterfactually depended on my *decision* not to jump in. For instance, we could imagine that I am very decisive type of person. So, had I not decided not to jump in, I *would* have decided to jump in and thus, I would have saved the child. That is, in the closest possible world where I don't decide not to jump in (a world where the neuroscientist failed, I am a decisive person, and there are no obstacles in the water), the

child lives. Hence, in FSOC thus conceived, the child's death counterfactually depends on my decision not to jump in. And I am responsible for this decision. Hence, if TR (Counterfactual) were true, it would follow that I am responsible for the child's death, and then, for my failure to save the child, in FSOC. So FSOC would be a counterexample to OA.

But TR (Counterfactual) *isn't* true, as shown by the following case. Imagine that a runaway train is going along the main track (track A) when it approaches a switch. I am by the switch and I see that, if the train continues on track A, it will run over a person, Victim, who is standing further down the track. Although switching to B wouldn't help, switching to C would, as depicted by the following picture:



Imagine that I have an irresistible urge for flipping switches and that I don't care whether Victim lives or dies. I think about where to flip the switch, and I decide to switch it to B. Victim dies. Unbeknownst to me, a neuroscientist has been monitoring my brain and is in control of my bodily movements. Had I decided to switch the train to C, he would have forced my finger into the B-position. So I couldn't have saved Victim.

In this case, given that I freely decided to flip the switch to B, I am responsible for flipping the switch to B (although I couldn't have done otherwise). But I am not likewise responsible for Victim's death. After all, the case has the same structure as Sharks in this respect: I could have decided to switch the train to C, but I couldn't have saved Victim, for the neuroscientist would have stopped me. (Compare: in Sharks, I could have decided to jump in, but I couldn't have saved the child, for the sharks

would have stopped me.)[39] But Victim's death counterfactually depends on my flipping the switch to B: had I not flipped it to B, I would have flipped it to C and Victim would have lived. For, in the closest possible world where I don't flip the switch to B (a world where the neuroscientist fails and where I still have an urge for flipping switches) I flip it to C and Victim lives. Thus, I am responsible for switching the train to B, I am not responsible for Victim's death, but Victim's death counterfactually depends on my switching the train to B. In other words, TR (Counterfactual) is false.[40]

At the same time, this case serves to illustrate the fact that there can be counterfactual dependence without causation.[41] In this case, Victim's death counterfactually depended on my switching the train to B but wasn't caused by it. If anything, Victim's death was caused by *my failure to switch the train to C,* the safe track. For the fact that I switched the train to the unsafe track is relevant to Victim's death *only to the extent that* it entails the fact that I didn't switch the train to the safe track. This suggests that the cause of Victim's death was my failure to switch the train to C, not my switching it to B instead.[42]

As a result, what emerges from the example is that, when causation is absent, responsibility does not transmit to the ensuing outcome, even if there is counterfactual dependence. Thus, even if I was responsible for switching the train to B, my responsibility does not transmit to Victim's death, because my switching the train to B didn't cause Victim's death. The example also supports the idea that, when there is causation, responsibility does transmit to the outcome (with the usual provisos). Thus, in the example, *had* I been responsible for failing to switch the train to C, then I would *also* have been

---

[39] Given that everybody agrees about Sharks, everybody should agree about this case. In particular, the Frankfurt-style objector to OA should. Recall that he wants to draw a difference between FSOC and Sharks. He wants to say that, if I couldn't have made the morally right decision (as in FSOC), then I am responsible for failing to save the child; by contrast, if I could have made the morally right decision but I would have been stopped *afterwards* (as in Sharks), then I am not. The train case is analogous to Sharks, not to FSOC: by assumption, the neuroscientist would have intervened *after* I made the decision to switch to C, by forcing me to switch to B.

[40] Notice that it was foreseeable that my flipping the switch to B would be followed by Victim's death. Hence, the failure of transmission of responsibility cannot be blamed on the usual epistemic provisos, which are met in this case.

[41] *Pace* some theories of causation that regard counterfactual dependence as sufficient for causation. See, e.g., Lewis (1986a).

[42] Notice the parallel with my discussion of FSOC in the preceding section.

responsible for Victim's death, given that my failing to switch to C caused Victim's death. In this case, however, I was not responsible for failing to switch the train to C, because I couldn't have switched it to C (again, the case is like Sharks in this respect: I could have decided to switch the train to C, but I couldn't have acted on that decision because the neuroscientist would have stopped me). Thus, I am not responsible for Victim's death.[43]

I have argued that the most plausible candidate for transmitting responsibility to outcomes is causation. I don't intend this to be a knockdown argument against any other possible candidate, but I think it does make a good prima facie case for my view. In a nutshell, the view that I have defended is the following:

> TR (Causal): An agent's responsibility for X transmits to an outcome Y iff X causes Y (and the usual provisos obtain).[44]

TR (Causal) says that causation is necessary for the transmission of responsibility to outcomes (other notions of dependence by themselves won't do), and it says that it is also sufficient, with the usual provisos.

In what follows, TR (Causal) will play a double role: it will generate the *right* reasons for rejecting OA, and it will support the new moral asymmetry between actions and omissions. I take these up in turn in the next two sections.

---

[43] Again, notice the parallel with my discussion of FSOC. In FSOC, *had* I been responsible for failing to decide to jump in, I would *also* have been responsible for the child's death, because my failing to decide to jump in caused my the child's death. However, as I pointed out in the last section, it is not clear that I was responsible for failing to decide to jump in (what *is* clear is that I was responsible for *deciding* not to jump in).

[44] J. Feinberg endorses a view of this type in Feinberg (1970).

4. The *right* reasons for rejecting OA

As we have seen, FSOC is not a counterexample to OA, or, at least, the Frankfurt-style objector hasn't given us good reason to believe that it is. I submit that, by contrast, the following case *is* a counterexample:

> *Planted Sharks*: This time *I* am responsible for the sharks being in the water: yesterday I negligently released the sharks in the area—I had no good reason to release the sharks, and I had good reason not to, but I still did. Today I see that the child is drowning but I do not attempt a rescue because it would be fruitless.[45]

I will argue that Planted Sharks is a counterexample to OA because, although I couldn't have saved the child today, I am responsible for failing to save him today. Thus, it is possible to be responsible for an omission when one couldn't have done otherwise, and so OA is false.

The reason that I couldn't have saved the child in Planted Sharks is, again, that the sharks are in the water.[46] And the reason that I am responsible for my failure to save the child is that I am responsible for planting the sharks, which was one of the causes of the child's death. This is to say, unlike what was the case with FSOC, an argument from dependence succeeds in showing that I am responsible for failing to save the child in Planted Sharks. The argument appeals to the right kind of dependence, namely, causation, and it goes as follows:

(7) In Planted Sharks, I am responsible for planting the sharks yesterday.

(8) My planting the sharks yesterday caused the child's death today.

(9) TR (Causal): An agent's responsibility for X transmits to an outcome Y iff X causes Y (and the usual provisos obtain).

(10) Therefore, I am responsible for the child's death today.

---

[45] See also Clarke's drunk driver example in Clarke (1994).

[46] Couldn't I have saved him by not planting the sharks yesterday? To dissipate any worries of this sort, imagine that someone else would have planted the sharks had I not done so myself. Then I couldn't have saved the child, not even by not planting the sharks. Still, given that I planted the sharks, I am responsible for failing to save the child today.

(11) If I am responsible for the child's death, then I am responsible for my failure to save the child.

(12) Therefore, in Planted Sharks, I am responsible for my failure to save the child today.

This argument seems to go through. I am clearly responsible for planting the sharks, and my planting the sharks is clearly a cause of the child's death. Thus, it follows by TR (Causal) that I am responsible for the child's death. And, if I am responsible for the child's death, I am presumably also responsible for my failure to save the child.[47]

In this section, I presented the *right* reasons for rejecting OA. I argued that Planted Sharks is a counterexample to OA, and I used TR (Causal) to support this claim. In the next section, I will argue that a *different* asymmetry thesis is unscathed by Planted Sharks. Again, this new asymmetry thesis will be supported, at least in part, by TR (Causal).

5. The new moral asymmetry

In the last section I pointed out that, in Planted Sharks, the child's death today was caused by my planting the sharks in the water yesterday. Presumably, however, it *wasn't* also caused by my failure to jump into the water today. After all, jumping in today wouldn't have helped. Otherwise put: in Planted Sharks, I brought about the child's death today by bringing about *yesterday* that I could not prevent that death today, *not* by not trying to prevent a death *today* that I couldn't have prevented. In Planted Sharks, then, the fact that the child's death would have occurred anyway deprives my omission to jump in of its causal powers with respect to the death.

[47] Two words of clarification. First, notice that the usual provisos of TR (Causal) are met in Planted Sharks. By assumption, I was guilty of negligence in planting the sharks. So, in particular, it was foreseeable that my planting the sharks was likely to be followed by a harm of the sort that ensued. Second, is step (11) warranted in this case? I think it is. Start by imagining that today there are *other* people around the child that cannot save him as a result of my having planted the sharks yesterday. Clearly, I would be responsible for these other people's failures to save the child in that case. Similarly, it seems to me that, if I am the only person standing close to the child today and if I negligently released the sharks yesterday, then I am responsible for *my own* failure to save the child today.

Now, this suggests that there might be a new asymmetry between actions and omissions lurking in the background. For actions don't generally behave in this way. Take my action of planting the sharks, for instance. Imagine that, unbeknownst to me, had I not planted the sharks, someone else would have (and there's nothing that I could have done to prevent it). Still, given that I planted the sharks, I caused the child's death. In other words, although the child's death would still have occurred had I not planted the sharks, my planting the sharks caused it. By contrast, as I have pointed out, given that the child's death would still have occurred even if I had jumped in, my omission to jump in *didn't* cause it.

The asymmetry between actions and omissions that seems to be lurking in the background is a *causal* asymmetry: it is an asymmetry in the conditions under which actions and omissions can be causes. It is expressed by the following claim, the *New Asymmetry (Causal)* claim:

> NA (Causal): An action can cause an outcome even if the outcome would still have occurred in the absence of the action. By contrast, an omission cannot cause an outcome if the outcome would still have occurred in the absence of the omission.

According to NA (Causal), an agent can cause an outcome by *acting* a certain way even it the outcome would still have occurred had he not acted by way, but an agent cannot cause an outcome by *failing to act* a certain way if the outcome would still have occurred had he acted that way.

Of course, a few examples aren't enough to support a general principle like NA (Causal), and, in particular, the general claim that it makes about omissions. I will argue for NA (Causal) in the next section. Now, however, I will show that, *on the assumption that NA (Causal) is true,* a new moral asymmetry emerges between actions and omissions. This new moral asymmetry is based on NA (Causal) and TR (Causal), the principle of transmission of responsibility that I defended in section 3.

Recall TR (Causal):

> TR (Causal): An agent's responsibility for X transmits to an outcome Y iff X causes Y (and the usual provisos obtain).

From TR (Causal) and NA (Causal), this claim follows:

> NA: An agent's responsibility for an action can transmit to an outcome even if the outcome would have occurred anyway in the absence of the action. However, an agent's responsibility for an omission cannot transmit to an outcome if the outcome would have occurred anyway in the absence of the omission.

NA is the *New Asymmetry* claim. It is a moral asymmetry between actions and omissions: an asymmetry in the conditions under which one can be responsible for outcomes *by* action and *by* omission (in virtue of having acted a certain way and in virtue of having failed to act a certain way). Alternatively put, it is an asymmetry in the conditions under which responsibility for actions and omissions *transmits* to outcomes.[48]

Take the example of Sharks, for instance. In Sharks, I am responsible for an *omission* of mine, failing to jump into the water (since, as far as I knew, I could have saved the child by jumping in). Now, given the presence of the sharks, the child's death would still have occurred had I jumped in. Thus NA entails that my responsibility for failing to jump in *doesn't* carry over to the child's death. This is the right result.[49] Contrast this with Planted Sharks. In Planted Sharks, too, the child's death would still have occurred had I jumped in. However, in contrast with Sharks, in Planted Sharks I am responsible for an *action* that I performed yesterday—namely, planting the sharks—in virtue of which I cannot save the child today. Thus, NA allows us to say that I am responsible for the child's death because I am

---

[48] It is important to distinguish NA from a different asymmetry that some people have defended, according to which there is a moral difference between causing a death by action and causing it by omission, or between killing someone and letting someone die. Some people have suggested that, other things being equal, killing is worse than letting die. But the discussion of whether this is true centers around cases where the outcome *wouldn't* have occurred anyway, in particular, it is generally assumed that one *lets* someone die by failing to act only if, by acting, one *would* have saved the person. Thus, the question whether killing is worse than letting die is not likely to shed any light on the question whether NA is true. For discussion of the killing and letting die distinction, see, e.g., Bennett (1994) and Foot (1994).

[49] What if, at the same time that I failed to jump into the water, someone else failed to put the sharks in a cage? (Imagine that it was this other person's job to do so.) Aren't we then jointly responsible for the child's death? I think so, but I think that our individual failures still don't cause the death. Rather, our *joint* failure causes the death (see chapter 3).

responsible for planting the sharks.[50] Notice that NA would allow us to say this even if it turned out that, had I not planted the sharks yesterday, someone else would have, and thus, even if it turned out that, had I not planted the sharks yesterday, the child's death would still have occurred today. Again, this is the right result.[51]

Here are another two scenarios that serve to illustrate the content of NA. In the first scenario, "Drunken Doctor," a tumor is found in a patient's brain, and the surgeon who is to operate on the patient gets drunk before the operation. During the operation, he fails to perform a cut that would have served to remove the tumor, and the patient dies as a result. Had the doctor not been drunk, however, he would have recognized the need for the cut, he would have made the cut and removed the tumor, and the patient would have lived. Clearly, the doctor is responsible for the death of his patient in Drunken Doctor. In the second scenario, "Lucky Doctor," a tumor is found in a patient's brain, and the doctor negligently fails to make the cut that doctors normally make in cases of that sort to try to remove the tumor. The patient dies from the tumor. However (unbeknownst to the doctor at the time of the operation), the tumor was too deep into the patient's brain and couldn't have been removed, so, as it turns out, making the cut wouldn't have helped (imagine that there is no way to remove tumors that deep without killing the patient). In this case, it seems that, although the doctor is responsible for not making the cut (any good surgeon would have made the cut in his place), he is not similarly responsible for the patient's death. In this respect, the doctor is morally lucky: as it turned out, the patient wouldn't have lived if he had made the cut, so he is not responsible for the death by virtue of not making the cut. Thus, Drunken Doctor and Lucky Doctor

---

[50] That is, I am responsible for the child's death, but the responsibility is inherited from my planting the sharks, not from my failing to jump in. It is inherited from the former and not from the latter because the former, not the latter, is a cause of the death.

[51] NA only *allows* us to say this, instead of *forcing* us to say this, because—unlike TR (Causal)—it doesn't provide sufficient conditions for the transmission of responsibility, but it is only concerned with a *necessary* condition for the transmission of responsibility that exists in the case of omissions but not in the case of actions.

differ with respect to the doctor's responsibility for the patient's death: while the drunken doctor is responsible for the death, the lucky doctor is not.[52]

What would NA say about these cases? About Lucky Doctor, NA would say that, even if the doctor is responsible for an omission of his, his not making the cut, the responsibility does not carry over to the patient's death. For the death would still have occurred if he had made the cut. This is the right result. Now consider Drunken Doctor. In Drunken Doctor, had the doctor made the cut, the patient *wouldn't* have died. Hence NA does *not* entail that the doctor's responsibility for his failure to make the cut doesn't carry over to the death. NA only says that an agent's responsibility for his omission does not carry over to an outcome when the outcome would *still* have occurred in the absence of the agent's omission. In cases like Drunken Doctor, however, the outcome would *not* have occurred in the absence of the agent's omission. Therefore, NA is consistent with the claim that the drunken doctor is responsible for the patient's death in virtue of his not making the cut. Again, this is the right result.[53]

In this section I motivated NA (Causal) by appeal to examples, and I explained how the new moral asymmetry claim, NA, follows from NA (Causal) and TR (Causal). I already argued for TR (Causal) in section 3. Thus, in order to complete my argument for NA, I must argue for NA (Causal). I turn to this in the next section.

6. The causal asymmetry

Recall NA (Causal):

> NA (Causal): An action can cause an outcome even if the outcome would still have occurred in the absence of the action. By contrast, an omission cannot cause an outcome if the outcome would still have occurred in the absence of the omission.

---

[52] I emphasize that we would probably still want to say that both doctors are *bad* doctors, maybe even *equally bad* doctors. I remind the reader that I am relying on the distinction between judgments about character and judgments about responsibility for outcomes (see n. 29).

[53] Again, NA is only *consistent* with this claim instead of *entailing* it because it doesn't provide sufficient conditions for the transmission of responsibility (see n. 51).

Why should we believe that there is such a difference between the causal powers of actions and omissions? In what follows, I offer a reason to believe this.

Start by contrasting the following two types of situations, which I will call *switches* and *non-switches*. An example of a *switch* is the following case:

> *Redirecting the Train*: A train is hurtling down a track, where a person, Victim, is standing (up ahead). There is a spur of track on the right, and a switch. I flip the switch and as a result the train turns; however, the two tracks reconverge before the spot where Victim is standing. Victim gets killed by the train.

By contrast, here is an example of a *non-switch*:

> *Pulling the Trigger*: Assassin is about to shoot Victim. I don't realize this, and since I want Victim to die too, I pull the trigger of my own gun. Victim dies as a result.

Here is yet another *non-switch:*

> *Removing the Shield*: Assassin is about to shoot Victim. There is a bulletproof shield that would have stopped the bullet before it reached Victim. I remove the shield. Assassin shoots and Victim dies as a result. Had I not removed the shield, however, Backup would have, and Victim would still have died.

Fill in the details of the cases so that in none of these cases there is something that I could have done to save Victim: Victim would still have died, had I done anything else. (Imagine, in the first case, that I cannot stop the train and that there are no other tracks to divert the train to; in the second case, that I cannot stop Assassin; and, in the third case, that I cannot stop either Assassin or Backup.) If so, in all of these cases, switches and non-switches, what I do does not affect the outcome—Victim would still have died, had I done anything else—but merely helps to determine the route by which the outcome occurs.

However, there is an important difference between switches and non-switches. Intuitively, I *don't* cause the death in a switch like Redirecting the Train, but I *do* cause the death in a non-switch like Pulling the Trigger or Removing the Shield. Intuitively, when one merely redirects a train that was already going to hit and kill a person, one *doesn't* thereby cause the person's death; by contrast, when one

shoots the bullet that kills a person, or when one removes an obstacle to an ongoing bullet, one *does* cause the person's death, even if the death would still have occurred had one not done those things.[54]

Now, what is the source of this causal difference between the two types of case? Intuitively, it seems to be the following. Even though in both types of case the outcome would still have occurred had I failed to act the way I did, in a *non-switch* the agent creates a new threat or promotes a preexisting threat. Thus, in Pulling the Trigger, the death would have occurred even if I hadn't shot, because Assassin would have shot, but by pulling the trigger I created a new threat that went to completion. Hence, I caused his death. And, in Removing the Shield, the death would have occurred even if I hadn't removed the shield, because Backup would have removed it, but by removing the shield I promoted the old threat to Victim: I removed an obstacle to it, and the threat then went to completion. Hence, I caused Victim's death. By contrast, in a *switch*, there is a preexisting threat and the agent does not promote it or create a new one. At most, the agent diverts the preexisting threat onto a slightly different path, as in Redirecting the Train. Hence, the agent does not cause the outcome by acting the way he does.

I will assume that this account of the causal difference between switches and non-switches is intuitive enough and can be put to work successfully. A more complete explanation of the difference should probably contain an account of the concepts of a threat, creation of a threat, and promotion of a threat. Unfortunately, I cannot do this here. Thus, I have to rely on intuition to ground my claims about the difference between switches and non-switches. But it does seem intuitive enough, first, that there is a causal difference between switches and non-switches and, second, that the causal difference is grounded in a difference between creating a new threat or promoting an old threat, on the one hand, and neither creating a new threat nor promoting an old threat, but at most redirecting an old threat, on the other.

Now, assuming this story about the causal difference between switches and non-switches, I suggest that a similar story can be told in support of NA (Causal). NA (Causal) says that the fact that an

---

[54] I stress that this is only an appeal to intuition. Some theories of causation don't yield the intuitive verdict about switches. For attempts at capturing our causal intuition about switches, see Rowe (1989) and Yablo (2002). In chapter 1, I argued that embracing the "difference-making idea" about causation (the idea that causes make a difference to their effects) requires respecting our causal intuitions about switches.

outcome was going to happen anyway deprives an agent's omission of its causal powers towards the outcome, while it doesn't necessarily deprive an agent's action of its causal powers towards the outcome. My suggestion is that this is so because every omission case where the outcome would have occurred anyway is a *switch,* but some action cases where the outcome would have occurred anyway are *non-switches*. We have already seen that some action cases where the outcome would have occurred anyway are non-switches (Pulling the Trigger and Removing the Shield are cases of this type). In order to see that NA (Causal) is true, then, it remains to be shown that any omission case where the outcome would have occurred anyway is a switch. This is what I do in what follows.

Consider Sharks again. Given the presence of the sharks in the water, I couldn't have prevented the death of the child by jumping in. Also, intuitively, by refraining from jumping in, I didn't create a new threat or promote the preexisting threat in any way, simply because I didn't *do* anything: I just let the ongoing process run its preset course. There was an existing process leading to the death, which was developing in a certain way, and, by failing to act, I let it keep developing that way. Intuitively, any time one doesn't interfere with a process, one doesn't promote it (or create a new one): one just lets it unfold. Thus, it seems that omissions don't promote existing threats or create new threats; at most, they let existing threats develop in certain ways. If so, any omission case where the outcome would have occurred anyway is a switch, in the way I have defined switches.

A word of clarification: I stress that I am *not* suggesting that, given that omissions don't promote existing threats or create new threats, they never cause anything. Many times, letting a process develop in a certain way *is* sufficient for causing an outcome. For instance, if a mother doesn't feed her baby when she could easily have fed him, and the baby starves to death, then the mother lets the process of starvation develop but she clearly causes the baby's death. Rather, what I am suggesting is that, *when the outcome would have occurred anyway,* an omission doesn't cause the outcome given that, first, the outcome would have occurred anyway and, second, omissions don't promote or create threats. This is what happens in Redirecting the Train, an ordinary switch: in Redirecting the Train, I don't cause Victim's death by

flipping the switch because, first, the death would still have occurred if I hadn't flipped the switch, and, second, by flipping the switch I don't promote the preexisting threat or create a new one.

Also, it is important to realize that, although NA (Causal) says that, when an outcome would still have occurred in the absence of an omission, *that omission* isn't a cause of the outcome, this is consistent with its being the case that something else that the agent did or failed to do caused the outcome. For instance, in Planted Sharks, my failure to jump in isn't a cause of the child's death, because the death would still have occurred if I had jumped in, but *my planting the sharks,* an action of mine, is a cause. Or imagine that I am an absentminded lifeguard. The child is drowning, and there are no sharks. Still, I wouldn't have saved the child by jumping in because, if I had jumped in, I would have forgotten the life preserver and thus I wouldn't have been able to save the child. According to NA (Causal), my failure to jump in isn't a cause of the child's death, because the death would still have occurred if I had jumped in. However, arguably, *my failure to jump in with the life preserver,* a *different* omission of mine, is a cause. (Presumably, my failure to jump in and my failure to jump in with the life preserver are different failures, since they obtain under different circumstances, and, of the two, the second has a better claim to being a cause.)

Now, all the cases that I discussed have a common structure. Their structure is the following. An agent fails to act in a certain way and an outcome occurs. Had the agent acted in the relevant way, the outcome would still have occurred because the agent's intervention wouldn't have sufficed to stop the ongoing threat, which would have still gone to completion. For instance, in Sharks, the drowning process is on its way, and I wouldn't have stopped it by jumping into the water: it would have gone to completion anyway. Call any case of this type an *easy* case. I think I have successfully argued that, in the easy cases, the agent's omission doesn't cause the outcome. However, there are cases of a different type, with a different underlying structure, which are more likely to make trouble for NA (Causal). I turn to these cases in the next section.

7. The hard cases and the issue of transitivity

A *hard* case is a case with the following structure. An agent fails to act in a certain way and an outcome occurs. Had the agent acted in the relevant way, the outcome would still have occurred; however, it would have occurred as a result of a *different* threat from the one that actually led to the outcome (call the threat that actually led to the outcome "the *actual* threat"). That is, the agent could have stopped the actual threat, but he still couldn't have prevented the outcome. He couldn't have prevented the outcome because there was a backup threat: had he stopped the actual threat, then the backup threat would have issued in the outcome just the same. Here is an example:

> *Evil Bystander*: Assassin is about to shoot Victim. I could stop him, but, since I want Victim dead too, I let him shoot. Victim dies. Unbeknownst to me, however, Backup is waiting in reserve. Had I stopped Assassin, then Backup would have shot, and Victim would still have died (imagine that I couldn't have stopped Backup, and imagine, also, that the details of the death would have been very similar, so it would have been the same death).

In Evil Bystander, I could have stopped the actual threat, but I still couldn't have prevented the outcome from happening because, had I acted, the backup threat would have issued in the outcome just the same.

Why is Evil Bystander a *hard* case for NA (Causal)? Because, in this case, one might be tempted to say the following. Given that I could have stopped the process that actually ended in Victim's death and I didn't, then, by failing to stop it, I contributed to the death. So my failure to stop Assassin was one of the causes of the death, although I couldn't have prevented the death. Hence, NA (Causal) is false: an omission can cause something that would have occurred anyway. In other words, the objection that a hard case like Evil Bystander can give rise to is that, contrary to what I have suggested, merely letting a threat unfold can be enough to cause an outcome, even if the outcome would have occurred anyway. In particular, the objection is that letting the *actual* threat unfold, when one could have stopped it, is enough to cause an outcome, even if the outcome would have occurred anyway via a backup threat.[55]

---

[55] J.J. Thomson believes that omissions can cause outcomes that would have occurred anyway in cases of this sort (Thomson (2003), p. 81). Her example is this. A prisoner is in a cell. Sally is supposed to give him water and Bert is supposed to give him bread, but neither complies. The prisoner dies from dehydration. Hence, Thomson claims, Sally's failure to give him water caused his death although the death would still have occurred (from starvation) if

In what follows, I argue that whether we think that we should revise NA (Causal) in light of hard cases like Evil Bystander essentially depends on our take on the transitivity of causation: we should revise it if we think that causation is transitive, but not so if we think that it is not transitive, or if we are unsure. I will briefly explain the main reasons for being skeptical about transitivity. However, I will not try to settle the issue of transitivity here, because it is an issue that deserves a much more extensive treatment than I can offer in this paper. As a result, I will leave it open whether NA (Causal) is in need of refinement, and I will briefly indicate how NA (Causal) *could* be revised in light of the hard cases, if one deemed it necessary.

Let us start by asking: why should we believe that, if one lets the actual threat unfold when one could have stopped it, one causes the outcome? The main motivation for believing this is the following. Since one could have stopped the actual threat but doesn't, one's failure to act is a cause of the actual threat's unfolding,[56] and the actual threat's unfolding is a cause of the outcome; hence, by transitivity, one's failure to act is a cause of the outcome. In Evil Bystander, my failure to stop Assassin was a cause of his shooting, and Assassin's shooting was a cause of Victim's death; hence, by transitivity, my failure to stop Assassin was a cause of Victim's death. In other words, the thought that, in the hard cases, the agent's omission causes the outcome because it lets the actual threat go to completion, is fueled by the thought that causation is transitive, which is expressed by the following principle:

Transitivity: If X causes Y, and Y causes Z, then X causes Z.

---

she had given him water. I prefer to focus on Evil Bystander for the following reasons. First, a person can survive a long time without food but not long without water. Hence, the death by starvation might not have been the same death, because it would have occurred at a much later time. Second, the death by starvation might not have been the same death because many of the details of the death would have been different. Third, even if the death by starvation would have been the same *death*, the *suffering* preceding the death would have been very different (since thirst is a lot more painful than hunger). Hence, there is the danger that our intuitions about the causes of the death get mixed up with our intuitions about the causes of the suffering preceding the death. Evil Bystander is specifically designed to sidestep these problems.

[56] This is the difference from the *easy* cases. In an easy case, I don't cause the actual threat's unfolding, since I couldn't have stopped the actual threat.

In the contemporary literature on causation, there is considerable debate over whether Transitivity is true. Some people believe that Transitivity is a general truth about causation, but other people think that it has counterexamples.[57] As I said, I will not try to settle this issue here; however, I will briefly explain what some of the problems with endorsing Transitivity are.

One of the problems with Transitivity is illustrated by Evil Bystander itself. As I have pointed out, Transitivity supports the claim that my failure to stop Assassin causes Victim's death in Evil Bystander: if Transitivity is true, then my failure to stop Assassin causes Victim's death because it causes Assassin's shooting, which causes Victim's death. However, by the same token, Transitivity supports the claim that, *had I stopped Assassin,* then *my stopping Assassin* would have caused Victim's death. For, had I stopped Assassin, then my stopping Assassin would have caused Backup to shoot, which would have caused Victim's death (recall that Backup is determined to shoot if Assassin doesn't). Hence, if Transitivity were true, it would follow that my stopping Assassin would have caused Victim's death too. In other words, I would have also caused the death if I had done the right thing. And this is counterintuitive.

Transitivity also yields puzzling results about Sharks. As we have seen, my failure to jump into the water in Sharks is not a cause of the child's death. But imagine that I *had* jumped in. Had I jumped in, then my jumping in would have caused the sharks to attack me, and the shark attack would have in turn caused the child's death[58]. Hence, if Transitivity were true, it would follow that my jumping into the water in Sharks would have caused the child's death. In other words, I *don't* cause the child's death by refraining from jumping in, but I *would* have caused his death by jumping in. This is an extremely counterintuitive result.

---

[57] For discussion of Transitivity, see, e.g., Hall (2000), Hitchcock (2001) and Paul (2000). In chapter 1, I argued that the difference-making idea enters in conflict with Transitivity, and that rejecting Transitivity in favor of the difference-making idea has the advantage that we can more easily account for our attributions of moral responsibility to agents.

[58] Just as, if I attack someone that is about to rescue the child, and then the child dies, I am a cause of the child's death.

I have explained how our take on the hard cases seems to heavily depend on our take on Transitivity, and I have explained what some of the main problems with Transitivity are. Now, could it be argued that the hard cases pose a problem for NA (Causal) *regardless* of what our take on Transitivity is? In particular, could it be argued that *intuition* alone shows that the hard cases are counterexamples to NA (Causal)?

I will argue that intuition alone is not enough to show this, because our intuitions about the hard cases are morally tainted. Take Evil Bystander. The way I set up the case, Evil Bystander is a case where my intentions are morally despicable. For I failed to stop Assassin when I thought that I could have prevented Victim's death by stopping him (since I was unaware of Backup's intentions). No wonder, then, that we feel tempted to say that I caused Victim's death. However, imagine that the details of the case are different. Imagine, for instance, that I am a good person and that I don't want for Victim to die. However, I still fail to stop Assassin. I fail to stop him because I see that Backup is waiting in reserve and I know that he will shoot if Assassin doesn't. Imagine, moreover, that I have good reason to let Assassin shoot instead of Backup: imagine, for instance, that I know that Backup is a very good shooter and I think that Victim will end up suffering more if I stop Assassin (say, because his hopes will be high again for a moment and then there will come the tragic realization that he will die anyway). In such circumstances, I find it a lot less plausible to say that I cause Victim's death by not stopping Assassin. After all, all I did was to do nothing, when there was nothing that I could have done to help.[59]

If so, this should be enough to cast doubt on our intuitions about Evil Bystander itself, where my intentions were bad. For none of the details of the sort I mentioned should be relevant to whether I caused Victim's death by failing to stop Assassin, since they only have to do with my intentions and beliefs (with what I thought was the case and with what I intended to be the case, but not with what was actually the case). Such details help to determine whether or not my intentions were morally good in

[59] I don't mean to suggest that it is impossible to say that I cause the death in this case (clearly, it would be possible to say that I cause the death but I am still not morally responsible for it). All I mean to suggest is that our *intuitions* about this case are different from those about Evil Bystander.

failing to stop Assassin, but not whether I caused Victim's death by failing to stop Assassin. This suggests that our intuitions about these cases are probably tainted by our moral judgments about the cases. Hence, we should probably not trust those intuitions.

I argued that, apart from Transitivity, there doesn't seem to be any compelling reason to believe that the hard cases undermine NA (Causal). Hence, unless we are moved by Transitivity, we shouldn't revise NA (Causal) in light of the hard cases. I also explained what some of the problems with Transitivity are. However, for all I have said, there might be powerful reasons to hold Transitivity despite some of its counterintuitive consequences. And, if so, NA (Causal) would be in need of refinement. In light of this, I will end my discussion of the causal asymmetry with a sketch of how NA (Causal) could be revised, if it were necessary to revise it.

If one were to revise NA (Causal) in light of the hard cases, then one should most likely proceed by restricting it to the conditions described in the *easy* cases. The easy cases, remember, are cases where the actual threat would have inevitably gone to completion, regardless of what the agent did. Thus, roughly, the revised causal asymmetry would be the claim that, *in circumstances where the actual process cannot be stopped,* omissions cannot be causes but actions still can. Sharks illustrates the claim about omissions: given the presence of the sharks, my jumping in wouldn't have stopped the ongoing drowning process, which would have gone to completion anyway; hence, my failure to jump in isn't a cause of the child's death. In turn, Removing the Shield (the example from the last section) illustrates the claim about actions: given the presence of Backup (who would have removed the shield had I not done so myself), my refraining from removing the shield wouldn't have stopped the ongoing shooting process started by Assassin, which would have gone to completion anyway; however, my removing the shield is a cause of Victim's death.

In other words, the revised causal asymmetry would be the claim that, when the actual threat cannot be stopped, an action can still cause the outcome because it can promote the threat, but an omission cannot cause the outcome because it cannot promote the threat. If one were to revise the causal asymmetry this way, one would then have to revise the *moral* asymmetry accordingly. Roughly, the

revised moral asymmetry would be the claim that, when the actual threat cannot be stopped, the responsibility for an omission cannot transmit to the outcome but the responsibility for an action can.

Let me sum up the results of the last two sections. I distinguished the easy cases from the hard cases. I argued that the easy cases clearly support NA (Causal). Then I explained why it might be thought that the hard cases make trouble for NA (Causal). Although I expressed some reservations about the hard cases, I argued that, ultimately, the decision whether to revise NA (Causal) in light of those cases is likely to depend on the answer to the question whether causation is transitive. Finally, I explained how NA (Causal) could be revised if causation turned out to be transitive.

8. Conclusions

In this paper, I argued for a new moral asymmetry between actions and omissions. I first explained my reasons for rejecting the old asymmetry thesis that some philosophers have suggested, and I then argued for a new way of understanding the moral asymmetry that the old asymmetry thesis attempted to rescue. On this new way of understanding the moral asymmetry, there is a moral asymmetry between actions and omissions because there is a more fundamental and purely *causal* asymmetry between actions and omissions, and because causation plays a fundamental role in the transmission of responsibility.

To conclude, let us examine the results of this chapter in connection with the results of chapter 1. In chapter 1, I argued for the following thesis:

> CDM (A/O): If an action caused an outcome, then the corresponding omission wouldn't have caused the outcome. Similarly, if an omission caused an outcome, then the corresponding action wouldn't have caused the outcome.[60]

CDM (A/O) is the result of restricting CDM to the case of actions and omissions of agents. In turn, in this chapter, I argued for the following thesis:

---

[60] Recall that this is short for the claim: "If an action caused an outcome, then, in the closest possible world where the action doesn't occur, the omission doesn't cause the outcome, and vice-versa: if an omission caused an outcome, then, in the closest possible world where the omission doesn't obtain (i.e., where an action of the relevant type obtains), the action doesn't cause the outcome."

> NA (Causal): An action can cause an outcome even if the outcome would still have occurred in the absence of the action. By contrast, an omission cannot cause an outcome if the outcome would still have occurred in the absence of the omission.

The most interesting claim that NA (Causal) makes is the claim about omissions. So let us focus on this claim:

> NA (Causal) [Omissions]: An omission cannot cause an outcome if the outcome would still have occurred in the absence of the omission.

Is there a connection between CDM (A/O) and NA (Causal) [Omissions]? In particular, does one entail the other?

It is clear that CDM (A/O) doesn't entail NA (Causal) [Omissions]. CDM (A/O) entails that an omission cannot cause an outcome if the corresponding action would have *caused* it. But from this it doesn't follow that, as NA (Causal) [Omissions] claims, an omission cannot cause an outcome if the outcome would still have occurred if the omission had been absent—i.e., if the corresponding action had *occurred.*

In what follows, I show that the converse entailment doesn't hold either: that NA (Causal) [Omissions] doesn't entail CDM (A/O). This is less obvious, however. For ease of reference, let us break up CDM (A/O) into two claims:

> CDM (A/O) [Claim 1]: If an action caused an outcome, then the corresponding omission wouldn't have caused the outcome.
>
> CDM (A/O) [Claim 2]: If an omission caused an outcome, then the corresponding action wouldn't have caused the outcome.

I will show that NA (Causal) [Omissions] entails Claim 2, but it *doesn't* entail Claim 1. As a result, NA (Causal) [Omissions] doesn't entail CDM (A/O).

NA (Causal) [Omissions] entails CDM (A/O) [Claim 2]:

(1) NA (Causal) [Omissions]: An omission cannot cause an outcome if the outcome would still have occurred in the absence of the omission.

(2) Suppose that an omission O caused an outcome U.

(3) Then U would not have occurred in the absence of O (that is, if the corresponding action A had occurred). (From 1 and 2)

(4) Then A would not have caused U. (From 3)

(5) Then CDM (A/O) [Claim 2] is true: if O caused U, A would not have caused U. (From 2-4)

NA (Causal) [Omissions] *doesn't* entail CDM (A/O) [Claim 1]:

To see this, let us try to build an argument from the former to the latter and let us then see where the argument fails:

(1) NA (Causal) [Omissions]: An omission cannot cause an outcome if the outcome would still have occurred in the absence of the omission.

(2) Suppose that an action A caused an outcome U, and let w be the closest possible world where A doesn't occur (where the corresponding omission O obtains).

(3) Then the counterfactual: "If O had not obtained, A would have caused U" is true in w. (From 2)

(4) Then the counterfactual: "If O had not obtained, U would still have occurred" is true in w. (From 3)

(5) Then O doesn't cause U in w. (From 1 and 4)

(6) Then CDM (A/O) [Claim 1] is true: if A caused U, O wouldn't have caused U. (From 2-5)

This argument is fallacious: 3 doesn't follow from 2. Suppose that A caused U in @ (the actual world). By assumption, w is the closest world to @ where O obtains. But the closest world to w where O doesn't obtain needn't be @ itself.[61] So the counterfactual "If O had not obtained, A would have caused U" needn't be true in w.

[61] S. Yablo draws attention to this feature of counterfactuals in Yablo (1992), p. 416.

Here is an example to illustrate. I water a plant in the actual world and the plant lives. However, unbeknownst to me, the state of the plant was very delicate, and it required exactly that much water to survive: less wouldn't have been enough and more would have been too much. Moreover, imagine that I generally pour a lot more water into my plants: this time I accidentally got distracted and I poured less than usual. Call the closest possible world to @ where I don't water the plant w. Now consider the world that is closest to w where I do water the plant. Presumably, this isn't going to be @, but a world in which I pour a lot more water. Call it w*. w* is a world where my watering the plant doesn't cause the plant to live, because the plant dies in that world. This shows that 2 doesn't entail 3: my watering the plant causes the plant to live in @, but it doesn't cause it to live in w*, the closest world to w where I water the plant.

We have seen that NA (Causal) [Omissions] doesn't entail CDM (A/O) [Claim 1]. Therefore, it doesn't entail CDM (A/O). Given that, as we have seen, the converse entailment doesn't hold either, it follows that the theses that I have defended in this and the preceding chapter are logically independent.

Chapter 3

How To Be Responsible For Something Without Causing It

1. Introduction

What is the relationship between being morally responsible[62] for something and causing it? Plainly, we are not responsible for everything that we cause. For we cause a multitude of things, including things that we couldn't possibly foresee we would cause and with respect to which it seems that we cannot be assessed morally. Thus, it is clear that causing something does not entail being responsible for it. But, doesn't the converse entailment hold? Doesn't being responsible for something entail causing it? Intuitively, it does: intuitively, we can only be responsible for things that we cause.

In this paper I will argue that this intuition is misguided. I will argue that we can be responsible for things that we don't cause, and thus being responsible for something does not entail causing it. Moreover, I will argue that this is so *for interesting reasons.* By this I mean two things.

First, I will argue that being responsible for something does not entail causing it, *even under the assumption that causation by omission is possible.* If, as some philosophers have argued, it were simply impossible to cause something by omission, then, clearly, responsibility would *not* require causation.[63] For there are things that we are responsible for not in virtue of what we do but in virtue of what we *fail* to do, i.e. in virtue of some of our omissions. So, if causation by omission were impossible, there would be things for which we are responsible without causing them. Following intuition, I will assume that

---

[62] In what follows, I will use the term "responsible" to mean *morally* responsible.

[63] For a recent defense of the view that there is no genuine causation by omission, see Dowe (2001). Note that even someone like Dowe, who denies the possibility of causation by omission, feels the pressure to bring the causal and moral realms back in-line to some degree. To this effect, he introduces the term "quasi-causation," and claims that responsibility requires causation *or* quasi-causation.

causation by omission *is* possible. Still, I will argue that being responsible for something does not require causing it.

Second, in order for the thesis that responsibility requires causation to be an interesting target, it must be properly restricted. For the unrestricted version faces a serious problem. The problem arises as follows. The unrestricted version says:

> *Moral Entails Causal (Unrestricted)*: For *any* X, if an agent is responsible for X, it is in virtue of the fact that he caused X, i.e. it is in virtue of the fact that one of his actions or omissions caused X.

The "in virtue of the fact that" is there to suggest that, when an agent is responsible for X, he is responsible *because* he caused X (although, presumably, this is only taken to be a *partial* explanation of why he is responsible: the fact that he caused X is a reason why he is responsible, but there are probably others).

Now, the following principle seems true:

> (P) In order to be responsible for something in virtue of having caused it, one has to be responsible for the cause itself.

P is plausible because, if one weren't responsible for the cause, then the fact that one caused the effect wouldn't have any tendency to explain why one is responsible for the effect.

The problem for Moral Entails Causal (Unrestricted) is that, in conjunction with P, it leads to an infinite regress. This emerges as follows. Suppose that we want to hold an agent responsible for an event X. Then, Moral Entails Causal (Unrestricted) says that one of the actions or omissions of the agent caused X. Call that action or omission "Y". Then, P says that the agent is responsible for Y.[64] But then,

---

[64] I am assuming that it makes sense to hold agents responsible for their own actions and omissions. This is, e.g., how J.M. Fischer and M. Ravizza use the concept of responsibility in Fischer and Ravizza (1998). See also van Inwagen (1978). As van Inwagen points out, we much more normally attribute responsibility to agents for events or states of affairs in the world than for their own actions or omissions. However, we do sometimes say things like "He cannot be held responsible for his actions, given that he was under the influence of that drug." In this paper, I will be assuming that the concept of responsibility applies to all of these categories of things.

Moral Entails Causal (Unrestricted) says that one of the actions or omissions of the agent caused Y. Call that action or omission "Z". And so on. This is a problem because it means that, in order for an agent to be responsible for something, he must be responsible for an *infinite* number of things. And, in principle, it is not easy to see how this could be so.

However, the thesis that responsibility requires causation can be restricted in a way that avoids this problem. We can, for instance, restrict it to outcomes in the external world (such as a person's death or a person's being harmed). The thesis that responsibility for *outcomes* requires causation is widespread among philosophers.[65] And it is no mystery why this is so. Clearly, we can only be responsible for what happens in the external world if we are hooked up to the world in some way. Now, the only way in which it seems that we could be hooked up to the world is by means of our actions and omissions. And the only way in which our actions and omissions could hook us up to the world seems to be by means of what they cause. Thus, the natural thought is that, if we are responsible for an outcome, it must be because our actions or omissions caused the outcome.

Moreover, we often seem to put this intuitive idea to work in the following way. Imagine that we believe that a person is responsible for a certain outcome. However, we then find out that nothing that the person did or failed to do caused the outcome. Then it is likely that we will abandon our belief that the person is responsible for the outcome. Imagine, for instance, that a sniper and I willingly fire our guns at the same time in my enemy's direction. My enemy dies and the autopsy reveals that only the sniper's bullet reached him and killed him. Then we will conclude that I am not responsible for my enemy's death, although I am responsible for trying to kill him (similarly, I will not be punished for his death, although I might be punished for a lesser crime, attempted murder). The reason why I am not responsible for the death is, intuitively, that I didn't cause it, even if I tried. In other words, the reason why I am not responsible for the death seems to be that my firing my gun, the only thing I did that could have made me responsible for the death, wasn't a cause of it.

---

[65] It is generally assumed that the only outcomes that agents can be responsible for are the causal consequences of their actions and omissions. See, e.g., Fischer and Ravizza (1998) and van Inwagen (1978).

In short, under the assumption that there is causation by omission, the following principle seems very plausible:

> *Moral Entails Causal*: If an agent is responsible for an outcome, it is in virtue of the fact that he caused it (some action or omission of his caused it).

In the first part of the paper I will argue that Moral Entails Causal is false. Then, in the second part of the paper, I will try to do some rebuilding. I will address the questions that the first part naturally gives rise to, namely: if we can be responsible for outcomes without causing them, then, does this mean that there is *no* connection between responsibility and causation? How can we be responsible for what goes on in the world without being causally connected to the world by our actions and omissions? I will make an alternative proposal about the relation between responsibility for outcomes and causation, and I will argue that the alternative proposal is, on reflection, as plausible and as helpful as Moral Entails Causal seemed to be.

2. The argument against the received view

Imagine the following situation. There was an accidental leak of a dangerous chemical at a high-risk chemical plant, which is on the verge of causing an explosion. The explosion will occur unless the room containing the chemical is immediately sealed. Suppose that sealing the room requires that two buttons—call them "A" and "B"—be depressed at the same time t (say, two seconds from now). You and I work at the plant, in different rooms, and we are in charge of accident prevention. Button A is in my room, and button B is in yours. We don't have time to get in touch with each other to find out what the other is going to do; however, we are both aware of what we are supposed to do. As it turns out, each of us independently decides to keep reading his magazine instead of depressing his button. The explosion ensues.

Now consider the following variant of the case. Again, button A is in my room, and I fail to depress it. This time, however, there is no one in the room containing button B. Instead, a safety

mechanism has been automatically set to depress B at *t*. When the time comes, however, B becomes stuck while it's up. Just as in the original case, then, neither button is depressed and the explosion occurs. Call the two cases "Two Buttons" and "Two Buttons-One Stuck," respectively. The cases differ in the respect that, in Two Buttons, B isn't depressed because you decided not to depress it, whereas, in Two Buttons-One Stuck, it isn't depressed because it got stuck.

I will argue that Two Buttons is a case of responsibility without causation, and thus it is a counterexample to Moral Entails Causal. My argument will take the following form:

(1) I am responsible for the explosion in Two Buttons.

(2) My failure to depress A didn't cause the explosion in Two Buttons-One Stuck.

(3) If my failure to depress A didn't cause the explosion in Two Buttons-One Stuck, then it didn't cause it in Two Buttons.

(4) Therefore, my failure to depress A didn't cause the explosion in Two Buttons. (From (2) and (3))

(5) No other action or omission of mine caused the explosion in Two Buttons.

(6) Therefore, Moral Entails Causal is false. (From (1), (4) and (5))

In other words, I will argue that I am responsible for the explosion in Two Buttons but nothing I did or failed to do caused it. In particular, my failure to depress A didn't cause it. And I will argue that my failure to depress A didn't cause the explosion in Two Buttons by arguing that it didn't cause it in Two Buttons-One Stuck and that my causal powers with respect to the explosion are the same in the two cases.

I take (5) to be clearly true.[66] In the following sections, I take up premises (1) through (3) in turn. First, however, a note on the methodology is in order. I will be arguing that we should accept the

[66] Intuitively, that is. On some views of causation, I caused the explosion by failing to do things that I couldn't possibly have done. In particular, on views according to which counterfactual dependence is sufficient for causation, I caused the explosion by failing to cast a spell that would magically prevent it (since, had I cast the spell, the explosion wouldn't have occurred). But recall that Moral Entails Causal says that, if an agent is responsible for an outcome, *it is in virtue of the fact that* some action or omission of his caused the outcome, that is, the fact that some action or omission of his caused the outcome *explains* why the agent is responsible for the outcome. Now,

premises in my argument because the price of rejecting them is very high. Now, it might well be that the price of giving up Moral Entails Causal, the thesis that my argument attacks, is *also* high (especially since, as I have granted, Moral Entails Causal is an intuitively plausible and fruitful principle). If so, we should do whatever comes at the *least* high price. And, how are we going to know what this is? Here is where the positive proposal of this paper steps in. I will argue that the price of giving up Moral Entails Causal is actually not high at all, for an alternative principle about the relation between causation and responsibility is at least as plausible and at least as fruitful. As a result, we shouldn't have qualms about abandoning Moral Entails Causal.

3. Argument for the first premise

In this section I will argue for premise (1):

(1) I am responsible for the explosion in Two Buttons.

I find (1) intuitively true.[67] In addition, there is a persuasive argument in support of this intuition. Briefly, the argument goes as follows. If we were to reject (1), then we would have to say that Two Buttons is a case of moral luck. But it would be wrong to count situations like Two Buttons as situations of moral luck. Hence, we should accept (1).

Let me explain. A case of *moral (good) luck* is a case where an agent, who behaves in a way that is generally conducive to a certain type of harm, is relieved of any responsibility for the harm (even if the harm ensues) thanks to the obtaining of some circumstances that are outside of the agent's control.[68] For

---

even if my failure to cast the spell were a cause of the explosion, this would have no tendency to explain why I am responsible for the explosion, since everybody else's failure to cast a similar spell would be a cause too. Hence, my point would stand in the end.

[67] And other people would too. A.M. Honoré writes: "If two huntsmen independently but simultaneously shoot and kill a third person, or two contractors independently fail to deliver essential building supplies on time, it is intuitively clear that each should be held responsible for the death or building delay." (Honoré (2002)) Two Buttons strikes me as analogous to Honoré's two contractors case.

[68] See Nagel (1979), ch. 3, and Williams (1981), ch. 2. There are also cases of moral *bad* luck, where an agent is responsible for an outcome partly due to circumstances that are out of his control.

instance, if I fire my gun at my enemy and the bullet is deflected by a gust of wind, but at the same time he is struck by a lightning and dies, then I am not responsible for my enemy's death, even if I acted badly. Thus, this is a case of moral luck. Now, if I weren't responsible for the explosion in Two Buttons, then Two Buttons would be a case of moral luck too. For it would be a case where I am not responsible for the ensuing harm even if I acted in a morally unacceptable way that is generally conducive to that type of harm. In addition, I would not be responsible for the harm due to the obtaining of some circumstances that were out of my control, namely, the fact that B also wasn't depressed.

However, I don't think that we are prepared to accept this as a genuine kind of moral luck. The first thing to notice is that, if we were to say that I am not responsible for the explosion in Two Buttons, then we would have to say the same about you. In other words, we would have to say that I am not responsible because B wasn't depressed, and you are not responsible because A wasn't depressed. Now, B wasn't depressed because *you* failed to depress it, and A wasn't depressed because *I* failed to depress it. Thus, if we said that Two Buttons is a case of moral luck, we would have to say that, for each of us, the fact that the other *also* behaved in a morally unacceptable way (that is generally conducive to a certain type of harm) is enough to relieve him of responsibility for the harm. But, in Two Buttons, the harm occurred in virtue of the fact that we behaved badly: had both of us done the right thing, the harm wouldn't have occurred. Thus, claiming that Two Buttons is a case of moral luck amounts to claiming that two wrongs that are generally conducive to a certain type of harm can neutralize each other in circumstances where they are jointly responsible for the occurrence of the harm. And this seems wrong.

In other words, we regard the situation in Two Buttons as one where a purely *human* failure took place, and thus we want to assign blame for what happened to the moral agents involved. The fact that the human failure is traceable to more than one human being does not mean that the agents are relieved of responsibility; rather, it means that they *share* responsibility for the bad outcome, just as the members of a gang share responsibility for a robbery.[69]

---

[69] This raises the question of how we should distribute the responsibility in cases of joint responsibility, which I will not try to answer here. Two Buttons is different from ordinary cases of joint responsibility in that what each of the

Contrast this with Two Buttons-One Stuck. In Two Buttons-One Stuck, B wasn't depressed due to a mechanical—not a human—failure. Intuitively, mechanical failures *are* the kind of thing that can give rise to moral luck. Intuitively, in Two Buttons-One Stuck, even though I thought that I could have prevented the explosion by depressing A, and even if I acted badly in failing to depress A, I was lucky. I was lucky because, at the moment at which I should have depressed A, B got stuck. The fact that B was stuck seems to exempt me from responsibility for the explosion. In other words, Two Buttons-One Stuck strikes us as a typical case of moral luck, where some natural phenomenon that is outside my control takes away my responsibility for the outcome.[70]

Two Buttons-One Stuck is similar to cases that philosophers have discussed in the context of the debate over whether one can be responsible for outcomes that one couldn't have prevented. Here is an example: I am walking by the beach when I see that a child is drowning. I think I could prevent his death, but I deliberately refrain from jumping in to attempt the rescue. The child drowns. Unbeknownst to me, however, there was a patrol of hungry sharks in the water that would have attacked me as soon as I jumped in, and hence I couldn't have saved the child. Am I responsible for the death of the child under those circumstances? It seems not. I am responsible for not trying to save the child, but not for his death.[71] Similarly, it seems that, in Two Buttons-One Stuck, I am responsible for not trying to prevent the explosion, but not for the explosion itself.

---

agents does is *sufficient* (independently of what the other agent does) for the outcome. But, if anything, this provides *more* reason (not less) to believe that each of the agents is responsible in Two Buttons.

Someone might want to try to say that Two Buttons is a case of collective responsibility without individual responsibility: a case where a group is responsible but none of the members of the group is. It has been argued that this phenomenon is possible. However, even if it were, Two Buttons is *not* likely to be an example. The type of case where—some people argue—there is collective responsibility without individual responsibility is one where the behavior of each of the agents is *excusable*. For instance, D. Cooper has argued that a society where some serious injustice has become an everyday practice, and where all of the living members of the society have been brought up to regard it as natural, is a case of collective responsibility without individual responsibility (see Cooper (1972), pp. 88-89). In Two Buttons, however, our individual behaviors *aren't* excusable.

[70] I think that there is moral luck of this sort, and I will be assuming that there is in the context of this paper. It is possible, however, to hold premise (1) without holding that there is moral luck (by claiming that I am responsible for the explosion both in Two Buttons and in Two Buttons-One Stuck).

[71] This is Fischer and Ravizza's "Sharks" case in Fischer and Ravizza (1998) (I discuss this case in chapter 2). A consequence of the fact that I am responsible for the explosion in Two Buttons is that I can be responsible for

In sum, there seems to be an interesting moral difference between Two Buttons and Two Buttons-One Stuck: it seems that, whereas I am responsible for the explosion in Two-Buttons, I am not responsible for the explosion in Two Buttons-One Stuck.[72] Now, what explains this difference? In both cases, I couldn't have prevented the explosion by depressing A, since the other button wasn't going to be depressed (in one case, because it got stuck; in the other case, because you failed to depress it). Why, then, am I responsible for the explosion in one case and not in the other? I will return to this question in section 7. As we will see in the next two sections, it is *not* because I cause the explosion in one case and not in the other. It will follow from my arguments in the next two sections that there is no causal difference between Two Buttons and Two Buttons-One Stuck.

4. Argument for the second premise

In this section I will argue for premise (2):

(2) My failure to depress A didn't cause the explosion in Two Buttons-One Stuck.

I will argue that we should endorse (2), or else we would be committed to much more causation by omission than we are prepared to accept.

Imagine that we said that my failure to depress A *did* cause the explosion in Two Buttons-One Stuck. To say that it did is to say that it caused it even if B was stuck and, hence, even if depressing A wouldn't have prevented the explosion. Had I depressed A *and had B not been stuck,* then the explosion

---

outcomes that I couldn't have prevented. (By contrast, Two Buttons-One Stuck and Sharks are cases where I am *not* responsible for an outcome that I couldn't have prevented.) How does my claim about Two Buttons square with what I say in chapter 2? It follows from NA (Causal) from chapter 2 that my failure to depress A isn't a cause of the explosion in Two Buttons, since it wouldn't have prevented it. I think that this is the right result, and I argue for this in the next section. Thus, my responsibility for the explosion isn't inherited from my failure to depress A. I will argue, however, that it is inherited from something else (see sections 6 and 7 below).

[72] *Pace* D. Parfit, who seems to believe that whether another moral agent is present or whether a natural mechanism is present cannot make a moral difference of the sort I am suggesting. See Parfit (1984), p. 82, and especially fn. 49. Parfit does not say why he thinks this. He might have been led to believe this after noticing that, as I will be claiming, whether another moral agent is present or whether a natural mechanism is present cannot make a causal difference, and then concluding that it cannot make a moral difference. This last step is the step I think we shouldn't take.

would have been prevented, but my depressing A wouldn't have been sufficient to prevent it. Still, it would have been a cause. Now, if this is so, we should probably say that an agent's failure to act in the relevant way caused the outcome in all of the following cases. If a child is drowning but there are sharks in the water that would have thwarted a rescue attempt, we would have to say that I caused the death of the child by failing to jump into the water to rescue him, when I couldn't have saved him, given the presence of the sharks. After all, had I jumped into the water to save him, *and had there been no sharks in the water,* I would have saved him. Similarly, we would have to say that a doctor's failing to operate on a patient with a tumor caused the patient's death, when he couldn't have saved him, for the tumor was too deep into the patient's brain and thus couldn't be removed. After all, had the doctor operated, *and had the tumor not been so deep into the brain,* the patient would have lived. Notice, in particular, that we would have to say this even if the doctor was fully aware of the fact that the tumor couldn't be removed (and, in the drowning case, even if I was fully aware of the presence of the sharks). For, presumably, an agent's epistemic state is irrelevant to the causal powers of his actions and omissions.

The recipe for obtaining implausible results of this type is easy to follow. Take any outcome that I couldn't have prevented by acting in a certain way, given the existence of an impeding factor. Still, had I acted that way, *and had the impeding factor been absent,* the outcome would have been prevented. Hence, any case of this type will share the structure of Two Buttons-One Stuck. Thus, if we said that I caused the explosion in Two Buttons-One Stuck, then we would probably have to say that I caused the outcome in all of these cases.

Moreover, whether there is one or *more than one* impeding factor couldn't possibly make a difference to my causal powers. Hence, the recipe is easily generalizable to cases where many impeding factors are in play. Here is an example. Suppose that a child is about to drown in the ocean. The only way to save him now would be by using a super fast boat. Suppose that I am the manufacturer of the only type of engine in town that could have been part of such a boat. Naturally, though, the engine wouldn't have been enough: many other materials would have been necessary to build the boat. But, had the engine, and the wood, and the steel, and the iron, and all the other things necessary to build the boat been

available yesterday, and had we put the parts together in the right way then, we could have saved the child by using the super fast boat today. If we rejected (2), then, we would have to say that my failure to provide the engine to build the boat caused the death of the child. But this is extremely implausible.

My claim, then, is that we should accept (2), or else we would be committed to far more causation by omission than it seems reasonable to accept. In addition, we would be committed to a lot of causation by omission *of a very peculiar type*. Let me explain. When we wonder what the causal history of a certain phenomenon is, we don't expect to find *redundant* causes of the phenomenon. This is to say, we expect every element in the causal history to play an indispensable role, which no other element in the causal history plays. Suppose, for instance, that we wish to reconstruct the causal history of the shattering of a window. If we find out that there were two rocks in the neighborhood of the window right before it broke, each of them with a momentum enough to break the window, then we will naturally regard the two rocks as competing for one and the same role in the causal history of the window's shattering. We will want to know, for example, if one of them deflected the other away from its path, or if one of them got to the window first and, in any of these cases, we will count one of them, but not the other, as a cause. The reconstruction of the causal history of the shattering would contain a redundancy if it made reference to both rocks. As a result, we will probably feel that, until we resolve the redundancy, we won't have given a precise reconstruction of the causal history of the shattering.

There might be exceptions. In particular, some people believe that there are phenomena (or, at least, there *could* be phenomena) whose causal history is essentially redundant. These cases are known as cases of *overdetermination*.[73] An example that is generally offered in the literature as an example of overdetermination is a window shattering case where two rocks impact the window at exactly the same time, with the same momentum, and where each of the rocks in isolation from the other would have been sufficient to break it, and in exactly the same way as it actually did. In that case, both rocks should probably enter into the reconstruction of the causal history of the shattering, even though each of them

[73] By "overdetermination" I mean what some philosophers have called "*symmetric* overdetermination."

renders the other redundant.[74] Now, although some people are inclined to accept the possibility of overdetermination, the overwhelming majority still regards it as a rare phenomenon. Indeed, it seems that we would be very surprised to find out that the causal histories of *many* events that occur are redundant in this way. However, I will argue that, rejecting (2) would have the implausible consequence that there is widespread overdetermination, and thus, that redundancy is everywhere.

If my failure to depress A caused the explosion in Two Buttons-One Stuck (this is to say, if (2) were false), then, by the same token, B's not being depressed would also be a cause. But each would have sufficed for the explosion to occur, and thus they would render each other redundant. In the fast boat case, the boat couldn't have been built unless I provided the fast engine and unless many different materials were available at the required time. Thus, if the absence of the fast engine and of each of those materials were all causes of the drowning of the child, the causal history of the drowning would be *multiply* redundant. But it is implausible to believe that redundancy is everywhere. Hence, this provides further reason to accept (2).[75]

---

[74] At least if we find out that the presence of one rock didn't have any effect on the contribution that the other rock actually had; otherwise, we might want to count them as joint causes. Imagine, for instance, that the rocks slightly touch each other right before they hit the window and this slows them down. As a result, the momentum of each of the rocks when they hit the window wouldn't have been independently sufficient to break the window. If so, the two rocks jointly caused the shattering.

[75] A famous objection to the view that the mental is causally efficacious even though it is not identical to the physical is that, if such a view were true, then there would be widespread overdetermination. (See Kim (1998), pp. 44-45.) A possible reply in the mental case is to say that the redundancy is only problematic when the redundant causes are *fully distinct*, and that mental states and physical states aren't fully distinct because the former supervene on the latter. Notice, however, that this kind of reply isn't available in the cases that matter to us. For instance, my failure to provide the fast engine and the absence of wood are fully distinct conditions.

Although most people agree that overdetermination is not widespread, there are a few exceptions. A notable exception is J. Schaffer. In Schaffer (forthcoming), he argues that overdetermination is everywhere. For instance, when a big rock hits a window flying northwards, the rock's eastern and western hemispheres (each of which would have been sufficient to break the window) overdetermine the window's shattering. Schaffer claims that these cases involve no improbable coincidences or conspiracies because the overdetermining parts are "lawfully yoked": natural forces hold the parts of the rock together. Thus, there is no good reason to think that overdetermination is improbable. Notice, however, that the overdetermination that we would be committed to in the boat case is not of this type: no natural force holds the different parts of the *unbuilt* boat together. And, even if Schaffer were right in that overdetermination by lawfully connected parts does not seem improbable, it is clear that overdetermination by lawfully *disconnected* parts *does* seem improbable.

5. Argument for the third premise

Finally, let us turn to the third premise in my argument:

(3) If my failure to depress A didn't cause the explosion in Two Buttons-One Stuck, then it didn't cause it in Two Buttons.

My argument for (3) will be based on the fact that the two cases, Two Buttons and Two Buttons-One Stuck, are relevantly similar. There are differences between them; as I have briefly indicated, they are enough to ground a moral difference between the cases. But I will argue that the differences that there are could not plausibly be viewed as mattering *causally*.

The cases have been laid out in such a way that the essential difference between them is that a *person* is in control of button B in one case, but a *mechanism* is in control in the other. In both cases, B isn't depressed at the required time, but, in Two Buttons, it is because you failed to depress it, whereas, in Two Buttons-One Stuck, it is due to a mechanical failure. Hence saying that (3) is false would amount to saying that whether there is a causal connection between my failure to depress A and the explosion can depend on whether *a person,* as opposed to an unconscious mechanism of some sort, is in the other room. But this is highly implausible. That is, it is highly implausible to believe that the mere fact that a person is in the other room, as opposed to a machine that behaves in relevantly similar respects as the person does, might make a difference to my causal powers. The fact that B wasn't depressed certainly matters to my causal powers, given that B's being depressed was necessary to prevent the explosion. But it seems that whether B wasn't depressed as a result of *a person's failing to depress it* or as a result of *a mechanical failure of some sort* simply shouldn't be relevant to whether I caused the explosion by failing to depress A.

Let me illustrate with an analogous example. In order for the example to be sufficiently analogous, I suggest that we look at a purported case of overdetermination. (The reason for choosing this type of example is that, as we saw in the last section, cases like Two Buttons-One Stuck have the basic structure of an overdetermination case in that, were we to say that there is causation, we would thereby

have to say that there is overdetermination). Take the case of the two rocks simultaneously hitting the window, and imagine that I threw one of the rocks. Now, suppose that we are trying to establish whether my throwing my rock was a cause of the shattering. Would it matter, for these purposes, whether the other rock was thrown by another person or by an unconscious mechanism (say, an automated catapult)? Clearly not. Whether a person or a catapult threw the other rock seems completely irrelevant to the causal powers of my throw. What does seem to matter is *whether* the other rock impacted the window (or whether it was just my rock that impacted it), and *how* it impacted it (in particular, whether it made any important difference to the shattering of the window, or whether my rock was responsible for the major crack that ended in its shattering). But whether all this happened because a sentient being or an unconscious mechanism threw the other rock is simply irrelevant to whether my throw caused the shattering.

Similarly, it seems that whether a person was in charge of B in the other room, or whether a mechanism was, should simply be irrelevant to whether my failure to depress A caused the explosion. What does seem relevant is whether B was depressed (if it had been depressed, then my failure to depress A would have been a cause), and what effect B's not being depressed had on the explosion (if depressing only A had been sufficient to prevent the explosion, then, again, my failure to depress A would have been a cause). But whether B wasn't depressed as a result of a human failure or as a result of a mechanical failure seems irrelevant to whether my failure to depress A was a cause of the explosion.[76]

I have argued that the differences between Two Buttons and Two Buttons-One Stuck do not matter causally. I would like to conclude my defense of (3) by drawing attention to the fact that the main

[76] Could we say that some *other* difference, which comes hand in hand with the difference between a person and a machine, is relevant to my causal powers in these cases? How about the difference in the *certainty* of the outcome? (The thought is: having a person in the other room makes it less certain whether B will be depressed than if there is a machine in the room that will inevitably fail.) My reply is twofold. First, in the two rocks case, whether the other rock was thrown as a result of a mere fluke or as a result of an extremely reliable process is not relevant to whether my rock caused the window to break. The only thing that matters is whether it was thrown, and in what way, not how likely its being thrown was at the time that it happened. Second, assume that the person in the other room is completely determined to do the wrong thing. This still doesn't get me off the hook. Thus, this strategy won't serve to account for the fact that I am responsible for the explosion in this case. As a result, we would still have to accept my conclusion, or reject a different premise in my argument.

existing theories of causation are likely to regard the two cases as causally on a par: they are likely to entail that, either my failure to depress A caused the explosion in both cases, or in neither case (although different theories might disagree about whether it is a cause in both cases or in neither case). To see this, let us quickly review the two main categories of theories of causation.

Traditionally, theories of causation are classified into "regularity" theories and "counterfactual" theories. Very roughly, and in its simplest version, a regularity theory deems something a cause when it is sufficient, in the circumstances, and given the laws, for the occurrence of the effect. And, also very roughly, and in its simplest version, a counterfactual theory deems something a cause when it is necessary for the effect (in the sense that, if it hadn't happened, the effect wouldn't have happened). A *regularity* theory is likely to entail that my failure to depress A is a cause of the explosion in *both* cases, Two Buttons and Two Buttons-One Stuck. For my failure to depress A was sufficient, in the circumstances, and given the laws, for the explosion. Whether there is a person or a mechanism in the other room is simply irrelevant to the fact that my failure to depress A was sufficient, in the circumstances and given the laws, for the explosion. A *counterfactual* theory, by contrast, is likely to entail that my failure to depress A is a cause in *neither* case. For, given that the explosion would only have been prevented by depressing both buttons, and given that the other button wasn't depressed, my failure to depress A was not necessary for the explosion (the explosion would have occurred even if I had depressed A). Again, whether the other button wasn't depressed due to a human or a mechanical failure is simply irrelevant to the fact that my failure to depress A was not necessary for the explosion.

Despite their differences, then, both regularity theories and counterfactual theories are likely to entail that the two cases are causally on a par. Naturally, there are many varieties of both regularity and counterfactual theories of causation, and I do not intend for this brief sketch of theories of causation to span them all. However, it does serve as an indication that the kinds of factors that philosophers have considered as causally relevant are not the kinds of factors that distinguish Two Buttons from Two Buttons-One Stuck. As a result, my claim that the two cases are causally on a par does not seem to be particularly controversial.

This concludes my discussion of the premises of my argument against Moral Entails Causal. To sum up, my argument has been the following:

(1) I am responsible for the explosion in Two Buttons.

(2) My failure to depress A didn't cause the explosion in Two Buttons-One Stuck.

(3) If my failure to depress A didn't cause the explosion in Two Buttons-One Stuck, then it didn't cause it in Two Buttons.

(4) Therefore, my failure to depress A didn't cause the explosion in Two Buttons. (From (2) and (3))

(5) No other action or omission of mine caused the explosion in Two Buttons.

(6) Therefore, Moral Entails Causal is false. (From (1), (4) and (5))

The following emerges from the discussion so far. Let us coin the phrase "causal luck" to refer to the following phenomenon: two factors that would have been causally efficacious if they had acted alone cancel each other out (causally) when they occur simultaneously.[77] Then what Two Buttons and Two Buttons-One Stuck show is that *causal* luck is more common than *moral* luck. At least in the case of omissions, causal luck obtains when, given that the two factors occur simultaneously, a certain abstract dependence relation that would otherwise have existed between each of the factors and the effect is absent.[78] Moral luck, by contrast, is sensitive to other features of the situation, in particular, it is sensitive to whether what breaks the dependence between each of the factors and the effect is another *moral agent*. As a result, causal luck can occur without moral luck. Hence, there can be responsibility without causation.[79]

---

[77] A mundane example is this: two rocks that would have broken a window in the absence of each other collide in midair and lose their momentum.

[78] More precisely, in this chapter I argued that this is so in at least a wide range of cases. In turn, in chapter 2, I argued for the *general* claim that, when an outcome doesn't counterfactually depend on an omission, the omission doesn't cause the outcome. Note that all the cases I discussed in this chapter are "easy" cases (for discussion of the easy cases and the difference from the hard cases, see section 7 of chapter 2).

[79] Although I cannot go into this here, I believe that there are also *action* cases where this divergence between causal luck and moral luck occurs. I think there are "mixed" cases too. In particular, some "preemptive

6. Towards the new view

What is the relationship between responsibility and causation, if it is not entailment? Before addressing this question, I will consider a different question that arises naturally in light of the preceding discussion. As we will see, this will also serve the further purpose of helping us rethink the relationship between responsibility and causation.

The question that arises in light of the preceding discussion is the following. If, as I have argued, my failure to depress A did not cause the explosion in Two Buttons (or in Two Buttons-One Stuck, but let us focus on Two Buttons) and, by similar reasoning, your failure to depress B didn't cause it either, then *what* did? Naturally, the chemicals having leaked out of the place where they were stored did, but this answer isn't fully satisfying: the buttons (in virtue of their not being depressed) seemed to have had something to do with how the explosion was brought about too. The explosion occurred (partly) *because* the buttons weren't depressed (because the preventive mechanism constituted by the pair of buttons wasn't activated). Hence, the reconstruction of the causal history of the explosion would seem incomplete if it didn't make any reference to the buttons whatsoever.[80] The question is, then, how should this gap in the causal history of the explosion be filled?

I will suggest that some *other* condition, not my failure to depress A, and not your failure to depress B, although one that is closely related to both, caused the explosion in Two Buttons. What is this other condition? To see what it is, consider first an example with just one agent. Imagine that an orchestra has delivered a wonderful performance. At the end of the concert, I am expected to clap. Instead, I remain completely still. As a result, Jim forms the belief that I was rude. What caused Jim's belief that I was rude? Clearly, it was my failure to clap. What is my failure to clap? It is my failure to

---

preemption" cases might have this feature. For discussion of preemptive preemption cases, see McDermott (1995) and Collins (2000).

[80] Recall that I am assuming that omissions can be causes. If so, the causal history of an explosion should contain the failure of the mechanisms that were set to prevent it as well as some "positive" causes.

simultaneously move my left hand and my right hand in particular ways. My failure to simultaneously move my left hand and my right hand in particular ways obtains just in case *either* I fail to move my left hand in particular ways at the required time *or* I fail to move my right hand in particular ways at the required time, or both. Had I moved just my left hand, I wouldn't have clapped, and thus Jim would still have thought that I was rude. Had I moved just my right hand, I wouldn't have clapped either, and thus Jim would have thought that I was rude too. Jim would only have failed to think that I was rude if I had moved *both* of my hands in the ways that clapping requires.[81]

Let us represent the different conditions schematically. Let F(L) be my failure to move my left hand (in the required way, at the required time) and let F(R) be my failure to move my right hand (in the required ways, at the required time). Then we may represent my failure to clap as $F(L \wedge R)$. $F(L \wedge R)$ is my failure to *both* move my left hand and my right hand (in the required way, at the required time), and it obtains whenever at least one of the individual failures obtains (i.e., it is equivalent to the disjunction of the individual failures, $F(L) \vee F(R)$). $F(L \wedge R)$ should be distinguished from each of the individual failures, F(L) and F(R), as well as from the condition that results from conjoining the two, $F(L) \wedge F(R)$, which obtains just in case *both* individual failures obtain. $F(L) \wedge F(R)$ entails both of F(L) and F(R), since every world where $F(L) \wedge F(R)$ obtains is a world where F(L) obtains and also a world where F(R) obtains. In turn, each of F(L) and F(R) entails $F(L \wedge R)$, since every world where F(L) obtains is a world where $F(L \wedge R)$ obtains and every world where F(R) obtains is a world where $F(L \wedge R)$ obtains.

---

[81] I am oversimplifying, since, presumably, I can clap by moving just one hand. However, we may assume that Jim would also have thought that I was rude if I had just moved one hand. If so, Jim's belief that I was rude seems to have been caused by my failure to simultaneously move my right hand and my left hand in particular ways. Also, one might worry that my failure to clap cannot be fully analyzed in terms of (and, in particular, as the disjunction of) my failure to move my left hand in certain ways and my failure to move my right hand in certain ways because clapping requires a certain *coordination* between the movements of the two hands, not just moving the hands in certain ways. We can avoid this problem by imagining that there is just one way in which I could have moved my left hand and my right hand that could have resulted in my clapping, or by imagining that Jim would still have thought that I was rude if I had moved my hands in any other way.

My claim, then, is that F(L ∧ R) caused Jim's belief that I was rude. F(L ∧ R) obtains in every world where I fail to move *at least one* hand. In all and only those worlds, Jim would have believed that I was rude. Thus, presumably, F(L ∧ R) is a cause of Jim's belief.[82]

I submit that the situation in Two Buttons is analogous. Two Buttons is a case with essentially the same structure as that of the clapping case, with the only difference that it involves two agents instead of one. In Two Buttons, the explosion would only have been prevented if we had simultaneously depressed A and B at *t*. As a matter of fact, we didn't simultaneously depress A and B (in particular, neither of us depressed his button). I submit that *our failure to simultaneously depress A and B at t* caused the explosion. What is our failure to simultaneously depress A and B at *t*? It is the condition that obtains just in case *either* I fail to depress A at *t*, *or* you fail to depress B at *t*, or both. This condition obtains in the actual world given that both of us failed to depress our buttons, but it also obtains in worlds where only one of us fails to depress his button.

If F(A) is my failure to depress A and F(B) is your failure to depress B, our failure to simultaneously depress A and B can be represented as F(A ∧ B). Just as in the clapping case, F(A ∧ B) should be distinguished from each of our individual failures, F(A) and F(B), as well as from F(A) ∧ F(B), the condition that obtains just in case *both* of us fail to depress our buttons. Instead, F(A ∧ B) obtains whenever *at least one* of us fails to depress his button (i.e., it is equivalent to F(A) ∨ F(B)). Also, just as in the clapping case, F(A ∧ B) is entailed by all of these conditions. And, finally, just as in the clapping case, there is good reason to believe that it is a cause of the outcome, the explosion. Why? For the same reason that my failure to clap is likely to be a cause of Jim's forming the belief that I was rude in the clapping case, namely, the fact that, given the circumstances, the explosion occurs in all and only the

[82] Doesn't this commit us to overly *disjunctive* causes? I think it is no worse (at least, not *much* worse) to believe that F(L ∧ R) can be a cause than to believe that any omission, e.g. F(L), can be a cause. F(L ∧ R) obtains just in case F(L) or F(R) obtains. Similarly, F(L) obtains just in case I don't move my left hand at all or I move it in either of a number of ways that are not constitutive of clapping.

worlds at which F(A ∧ B) obtains. These include worlds where both of us fail to depress our buttons, but also worlds where only one of us does.

Let me sum up. The question that I wanted to address in this section was this: If, as my argument against Moral Entails Causal suggests, my failure to depress A did not cause the explosion in Two Buttons, and the same goes for your failure to depress B, then what did? Certainly, the two buttons had something to do with the explosion's coming about. My reply is that our failure to simultaneously depress A and B did. This is a condition that obtains whenever at least one of us fails to depress his button.

The question will surely arise: Is it really possible to hold, as I am suggesting, that our failure to simultaneously depress A and B caused the explosion, but neither of our individual failures, which entail it, did? More generally, is the following scenario really possible: X entails Y, Y causes E, but X *doesn't* cause E? Yes, there are clear cases of this sort. Here is one. Little Suzy just learned that people are mortal. In particular, she just learned that Grandpa, who she adores, is going to die someday, and this made her cry. Now, imagine that, unbeknownst to Suzy, Grandpa just died of a heart attack. The fact that he died entails the fact that he was mortal. But the fact that Grandpa died didn't cause Suzy to cry: she didn't cry because Grandpa died (she doesn't even know that he died), but because he is mortal. So it is possible for X to entail Y, for Y to cause E, but for X *not* to cause E.

Now, how is this going to help us figure out the relationship between responsibility and causation? I turn to this question in the following section.

7. Causation as the vehicle of transmission of responsibility

How can I be responsible for the explosion in Two Buttons, if my failure to depress A didn't cause it? I suggest the following. In Two Buttons, I am responsible for the explosion, not in virtue of the fact that my failure to depress A caused it (since it didn't), but in virtue of the fact that *something for which I am responsible* caused it. In other words, I am responsible for the explosion because there is a cause of the

explosion that I am responsible for. What is this cause of the explosion that I am responsible for? I submit it's our failure to simultaneously depress A and B. In the last section, I argued that this condition is a cause of the explosion in Two Buttons. Now I will argue that I am responsible for it.[83] It will then follow that something for which I am responsible caused the explosion in Two Buttons.[84]

The reason why I think we should say that I am responsible for our failure to simultaneously depress A and B in Two Buttons is similar to the reason why we should say that I am responsible for the explosion. As we have seen in section 3, we should say that I am responsible for the explosion in Two Buttons (and so are you) or else there would be things that no one is responsible for but that depend exclusively on the morally unacceptable behavior of some moral agents. And this is implausible. Now, our failure to simultaneously depress A and B depends exclusively on the morally unacceptable behavior of some moral agents, namely, you and me. Hence we should say that I am responsible for our failure to simultaneously depress A and B, and so are you. In other words, we should say that each of us is responsible for this failure. If so, I am responsible for a cause of the explosion in Two Buttons.

By contrast, I am presumably responsible for none of the causes of the explosion in Two Buttons-One Stuck. First, as we have seen, my failure to depress A didn't cause the explosion. Hence, even if I am responsible for my failure to depress A, such failure wasn't a cause of the explosion. And, second, just as I am not responsible for the explosion given that B was stuck (as I have pointed out in section 3), presumably, I am also not responsible for the failure of the two buttons to be simultaneously depressed. Figuring out precisely *why* this is so would require an in-depth investigation of the intriguing phenomenon of moral luck, a task that I cannot pursue here. But the fact remains that B wasn't depressed due to a mechanical failure, not a human failure, and somehow this seems to take away my responsibility

---

[83] Recall that "responsible" means at least *partly* responsible. What I will argue is that, just as you and I *share* responsibility for the explosion in Two Buttons, we *share* responsibility for our failure to simultaneously depress A and B.

[84] Why don't I just say the following: I am responsible for the explosion in Two Buttons in virtue of the fact that something I failed to do *entails* a cause of the explosion? Because, as we will see, I am hoping that the new principle will help to account for the moral difference between Two Buttons and Two Buttons-One Stuck. But my failure to depress A also entails a cause of the explosion in Two Buttons-One Stuck.

for the fact that the two buttons weren't simultaneously depressed. Hence, even if the failure of the two buttons to be simultaneously depressed caused the explosion (as we saw in the last section), I am not responsible for that failure. It seems, then, that I am not responsible for any of the causes of the explosion in Two Buttons-One Stuck.

In sum, I suggest the following principle about the relationship between causation and responsibility (a principle that helps to explain the moral difference between Two Buttons and Two Buttons-One Stuck):

> *Causal Transmits Moral*: If an agent is responsible for an outcome, then it is in virtue of the fact that the agent is responsible for something that caused the outcome.[85]

In other words, according to Causal Transmits Moral, if I am responsible for an outcome, then it doesn't have to be the case that something that I did or failed to do caused the outcome, but it does have to be the case that I am responsible for one of the outcome's causes. The outcome's cause that I am responsible for might be an action or omission of mine, but it can also be, as in Two Buttons, the collective behavior of a group of agents.

Note that Causal Transmits Moral is restricted to *outcomes*, just as Moral Entails Causal was. This prevents Causal Transmits Moral from leading to an infinite regress (if it weren't thus restricted, then it would follow that, in order for an agent to be responsible for something, he would have to be responsible for one of its causes, and for one of the cause's causes, and so on). Given that it is restricted to outcomes, Causal Transmits Moral does *not* entail that one must be responsible for an infinite number of causes of an outcome in order to be responsible for the outcome.[86]

---

[85] Just as in Moral Entails Causal, the causal connection is also intended to be a *partial* explanation of the fact the agent is responsible for the outcome. Presumably, other conditions are required, such as, roughly, the fact that the agent could foresee that something that he is responsible for would likely cause the outcome.

[86] Just as my individual failure to depress A isn't an outcome, our joint failure to simultaneously depress A and B in Two Buttons isn't an outcome. Hence, Causal Transmits Moral doesn't say anything about it. For all Causal Transmits Moral says, I might be responsible for our joint failure without being responsible for any of its causes. How am I responsible for our joint failure, then? As I pointed out, an answer to this question will have to rely on an account of the conditions under which there is moral luck and the conditions under which there is no moral luck.

Independently of the two cases that have occupied us, Two Buttons and Two Buttons-One Stuck, Causal Transmits Moral is an intuitively plausible principle about the relation between responsibility for outcomes and causation. On the face of it, we can only be responsible for an outcome if we are responsible for one of its causes. For instance, it seems that I wouldn't be responsible for a death by shooting unless I were responsible for the bullet's piercing the person's heart, or for some other contributing cause, such as the fact that the person was standing in front of the gun.

Moreover, not only does Causal Transmits Moral seem plausible, but it also appears to be as fruitful as Moral Entails Causal seemed to be. The paper started out with the remark that Moral Entails Causal appears to explain my lack of responsibility in cases like the following: I shoot at my enemy and miss; however, at the same time, a sniper shoots the bullet that kills him. Intuitively, I am not responsible for my enemy's death, and Moral Entails Causal seemed to explain why: I could only be responsible for his death in virtue of having shot a bullet at him, but my bullet didn't cause his death; hence, I am not responsible for his death. I have argued that Moral Entails Causal is false. However, its substitute Causal Transmits Moral could explain equally well why I am not responsible in this case. Given that I am not responsible for any of the causes of the death (e.g., the sniper's shooting, or my enemy's standing within the sniper's shooting range), it then follows from Causal Transmits Moral that I am not responsible for the death.

We have seen that Causal Transmits Moral is an initially plausible way of understanding the relation between responsibility for outcomes and causation. In addition, it is an improvement over Moral Entails Causal in that it does not overlook cases like Two Buttons, where, arguably, an agent is responsible for an outcome without causing it. Finally, it is as useful as Moral Entails Causal seemed to be in that it successfully accounts for the lack of responsibility of an agent for an outcome in cases where he is not responsible for any of its causes. I conclude that there is good reason to believe that Causal Transmits Moral successfully captures the relation between responsibility and causation.

8. Conclusions

What lessons should we draw from all this? One lesson we should draw is the following. Agents are responsible for what happens in the external world in virtue of how they interact with it—by means of their actions and omissions. *However*, their actions and omissions can make them responsible for things by virtue of more than their causal powers: they can make agents responsible for things by virtue of the causal powers of larger collective behaviors that those actions and omissions are part of.[87]

Another lesson we should draw concerns the general way in which to regard the concepts of responsibility and causation in relation to each other. There is a strong temptation to regard causation as a condition on responsibility. Quite generally, being responsible for an outcome tends to be associated with, roughly, intentionally (or negligently) causing the outcome.[88] If, as I have claimed, Moral Entails Causal is false, then this is a mistake: one can be responsible for an outcome without even causing the outcome. It follows that we shouldn't view the relation between responsibility and causation as a kind of *entailment* relation. How should we view it, then? If, as I have argued, Causal Transmits Moral is true, then causation should be rather viewed as the *vehicle of transmission* of responsibility. This is to say, in order for an agent to be responsible for an outcome, there must be a causal link between an earlier thing (event, state of affairs, etc.) and the outcome along which the responsibility of the agent was transmitted.

To conclude, let us examine the results of this chapter in connection with the results of the previous chapters.

In the previous two chapters, the relation between causation and responsibility was an important focus of discussion. In chapter 1 (section 6), I appealed to the connection between causation and responsibility for outcomes to show that CDM helps explain the lack of responsibility of agents in some

---

[87] In the jargon of collective action and collective responsibility, it is possible to have *individual responsibility* without *individual causation*, as long as there is a specific kind of *collective causation*.

[88] For instance, J. Feinberg analyzes "being responsible for a harm" as having acted or failed to act in such a way that (a) one could foresee (or should have foreseen) that the action or omission was likely to lead to the harm, (b) the action or omission caused the harm, and (c) the causal connection between the action or omission and the harm was not deviant. (Feinberg (1970))

cases of moral luck (in particular, in cases like Switch-with-Main-Track-Unexpectedly-Connected). I claimed that, in such cases of moral luck, CDM supports the claim that the agent's action doesn't cause the ensuing outcome, and this, in turn, explains why the agent isn't responsible for the outcome. Now, what warrants this second inference? Why does the fact that the agent's action doesn't cause the outcome explain why the agent isn't responsible for the outcome? In light of the results of the present chapter, we can see that it would be a mistake to say that what warrants that inference is the principle according to which being responsible for an outcome requires causing the outcome, for this principle is false. That is, it would be a mistake to think that it follows from the fact that the agent's action didn't cause the outcome (and from the fact that nothing else the agent did or failed to do caused the outcome) that the agent isn't responsible for the outcome. However, the principle that I defended in this chapter—the principle that causation is required for the *transmission* of responsibility to outcomes—can explain equally well why the agent isn't responsible for the outcome in such cases of moral luck, on the basis that the agent's action doesn't cause the outcome. For, if the agent's action doesn't cause the outcome, then (given how the cases are set up) there is no cause of the outcome that the agent is responsible for in those cases, and thus, there is nothing from which the responsibility for the outcome can be inherited. As a result, the principle that causation is required for the transmission of responsibility to outcomes helps explain the fact that the agent isn't responsible for the outcome in those cases.

In chapter 2, I explicitly discussed the relation between causation and responsibility. As a result, some of the results of that chapter are importantly connected to the results of the present chapter. In particular, in section 3 of chapter 2 I argued for the following principle:

> TR (Causal): An agent's responsibility for X transmits to an outcome Y iff X causes Y (and the usual provisos obtain).

Where "the usual provisos" included, for instance, that the agent could foresee (or should have foreseen) that X was likely to be followed by Y (or an outcome of Y's type). I wasn't particularly concerned with laying out the provisos that, together with causation, are (jointly) *sufficient* for the transmission of

responsibility to outcomes, for the cases I focused on in that chapter were ones were, intuitively, those provisos were met. Rather, my arguments in that chapter focused on the claim that makes causation a *necessary* condition for the transmission of responsibility to outcomes, which follows from TR (Causal):

> TR (Causal) [Necessity Claim]: An agent's responsibility for X transmits to an outcome Y *only if* X causes Y.

In turn, the principle that I argued for in this chapter is the following:

> *Causal Transmits Moral*: If an agent is responsible for an outcome, it is in virtue of the fact that the agent is responsible for something that caused the outcome.

Under the assumption that the responsibility for outcomes is inherited from other things, these principles amount to the same thing. As a result, the view of the relation between responsibility for outcomes and causation that I argued for in the last two chapters is essentially the same view. However, each of the chapters is concerned with securing different aspects of that view. In chapter 2, I was particularly concerned with showing that no notion of dependence has the same ties to responsibility that causation has (e.g., counterfactual dependence doesn't). In this chapter, by contrast, I was particularly concerned with clarifying the form that these ties between causation and responsibility take. This is to say, although both arguments are arguments for the view that causation is required for the transmission of responsibility to outcomes, they emphasize and defend different aspects of that view. While one attempts to establish that *causation* is required for the transmission of responsibility to outcomes (no other notion of dependence is), the other attempts to establish that causation is required for the *transmission* of responsibility to outcomes (not for the *existence* of responsibility for outcomes itself). As a result, the arguments in the two chapters complement each other.

References

Bennett, J. (1994) "Whatever the Consequences," reprinted in B. Steinbrock and A. Norcross (eds.), *Killing and Letting Die,* New York: Fordham U.P., pp. 167-91.

Clarke, R. (1994) "Ability and Responsibility for Omissions," *Philosophical Studies* 73, pp. 195-208.

Collins, J. (2000) "Preemptive Preemption," *Journal of Philosophy* 97, 4, pp. 223-34.

Cooper, D. (1972) "Responsibility and the 'System'," in P.A. French (ed.), *Individual and Collective Responsibility,* Cambridge, MA: Schenkman.

Dowe, P. (2001) "A Counterfactual Theory of Prevention and 'Causation' by Omission," *Australasian Journal of Philosophy* 79, 2, pp. 216-26.

Feinberg, J. (1970) "Sua Culpa," in *Doing and Deserving: Essays in the Theory of Responsibility,* Princeton: Princeton U.P, pp. 187-221.

Fischer, J. M. and M. Ravizza (1998) *Responsibility and Control,* Cambridge: Cambridge U.P.

Foot, P. (1994) "Killing and Letting Die," reprinted in B. Steinbrock and A. Norcross (eds.), *Killing and Letting Die,* New York: Fordham U.P., pp. 280-89.

Frankfurt, H. (1969), "Alternate Possibilities and Moral Responsibility," *Journal of Philosophy* 66, pp. 829-39.

Hall, N. (2000), "Causation and the Price of Transitivity," *Journal of Philosophy* 97, 4, pp. 198-222.

Hall, N. (forthcoming), "Two Concepts of Causation," in J. Collins, N. Hall and L. Paul (eds.), *Causation and Conditionals,* Cambridge: M.I.T. Press.

Hitchcock, C. (2001) "The Intransitivity of Causation Revealed in Equations and Graphs," *Journal of Philosophy* 98, 6, pp. 273-99.

Honoré, A. M. (2002) "Causation in the Law", in E.N. Zalta (ed.), *Stanford Encyclopedia of Philosophy,* Summer 2002 ed., URL: http://plato.stanford.edu/archives/sum2002/entries/causation-law/.

Kim, J. (1998) *Mind in a Physical World,* Cambridge, MA: The M.I.T. Press.

Lewis, D. (1986a) "Causation," reprinted in *Philosophical Papers,* vol. II, New York: Oxford U.P, pp. 159-213.

Lewis, D. (1986b) "Counterfactual Dependence and Time's Arrow," reprinted in *Philosophical Papers,* vol. II, New York: Oxford U.P., pp. 32-66.

Lewis, D. (2000) "Causation as Influence," *Journal of Philosophy* 97, 4, pp. 182-97.

Mackie, J. (1993) "Causes and Conditions," reprinted in E. Sosa and M. Tooley (eds.), *Causation,* New York: Oxford U.P., pp. 33-55.

McDermott, M. (1995) "Redundant Causation," *British Journal for the Philosophy of Science* 46, pp. 523-44.

McGrath, S. (ms) *Causation in Metaphysics and Moral Theory,* M.I.T. Ph.D. thesis, 2002.

McIntyre, A. (1994) "Compatibilists Could Have Done Otherwise," *Philosophical Review* 103, 3, pp. 458-88.

Nagel, T. (1979) *Mortal Questions,* New York: Cambridge U.P.

Parfit, D. (1984) *Reasons and Persons,* New York: Oxford U.P.

Paul, L. (2000) "Aspect Causation," *Journal of Philosophy* 97, 4, pp. 235-56.

Rowe, W. (1989) "Causing and Being Responsible for What is Inevitable," *American Philosophical Quarterly* 26, 2, pp. 153-9.

Schaffer, J. (2000) "Causation by Disconnection," *Philosophy of Science* 67, pp. 285-300.

Schaffer, J. (forthcoming) "Overdetermining Causes," *Philosophical Studies.*

Thomson, J.J. (2003) "Causation: Omissions," *Philosophy and Phenomenological Research* 66, 1, pp. 81-103.

van Inwagen, P. (1978) "Ability and Responsibility," *Philosophical Review* 87, pp. 201-24.

Weinryb, E. (1980) "Omissions and Responsibility," *The Philosophical Quarterly* 30, 118, pp. 1-18.

Williams, B. (1981) *Moral Luck,* Cambridge: Cambridge U.P.

Yablo, S. (1992) "Cause and Essence," *Synthese* 93, pp. 403-49.

Yablo, S. (2002) "De Facto Dependence," *Journal of Philosophy,* pp.130-48.

Yablo, S. (forthcoming) "Advertisement for a Sketch of an Outline of a Proto-Theory of Causation," in J. Collins, N. Hall and L. Paul (eds.), *Causation and Conditionals,* Cambridge: M.I.T. Press.