Causal Modelling and Frankfurt Cases

SANDER BECKERS
Cornell University

Almost half a century after Frankfurt presented his famous challenge to the Principle of Alternative Possibilities, it is still unclear whether he was successful. There have evolved two camps on this issue, namely those who argue that the principle is an intrinsic feature of moral responsibility, and those who construct complicated examples to counter every such argument presented. Throughout this debate one encounters causal judgments at every turn. In fact, such judgments play a fundamental role in the arguments offered by both sides. In this paper we take advantage of the significant progress that has been made in the literature on actual causation, by reframing the debate over Frankfurt cases into a causal context. Instead of focussing on metaphysical assumptions regarding free will, we restrict ourselves to the task of expressing the examples that drive the debate using causal models. In doing so it becomes clear that much of the debate is driven by confusion regarding causal judgments, which should therefore be made explicit before any progress is possible. Further, we suggest that a moderate version of the Principle of Alternative Possibilities is entirely consistent with even the most advanced Frankfurt cases found in the literature.

1. Introduction

In 1969, Harry Frankfurt wrote an article that would prove to be a turning point in the literature on *moral responsibility* (Frankfurt 1969). In 1973, David Lewis wrote an article that would prove to be a turning point in the literature on *actual causation* (Lewis 1973). The former introduced a certain style of examples – *Frankfurt cases* – that aim to falsify a principle regarding responsibility that was until then accepted almost unanimously, namely the Principle of Alternative Possibilities – *PAP*. The latter introduced a definition of actual causation that aims to overcome the severe limitations of defining causation as *counterfactual depen-*

Contact: Sander Beckers <Sander.Beckers@cornell.edu>

https://doi.org/---

dence, while still remaining faithful to the widely accepted view that causation and counterfactual dependence are closely related. The former has set the stage for a debate between those defending the PAP, and those constructing Frankfurt cases in order to falsify it. The latter has set the stage for the development of definitions of actual causation that are likewise based on counterfactual dependence, but differ in the way that they handle paradigmatic examples.

However, there is one big difference between how each of these articles has influenced their respective fields. Although there has been an increase in the complexity and level of detail in the positions that are being put forward in the debate on Frankfurt cases, neither side is ready to concede and the debate appears to be pretty much stuck in a stalemate (Fischer 1994; Timpe 2006). The counterfactual tradition on actual causation, on the other hand, has made significant progress in producing definitions of causation that improve upon the definition set forth by Lewis, which is now unanimously considered as an interesting but flawed first attempt (Woodward 2003; Paul & Hall 2013; Weslake 2015; Halpern 2016; Beckers & Vennekens 2017b). (Even Lewis (2000) himself abandoned his original definition, in favour of a more nuanced proposal.)

Besides offering a case study in the historical analysis of philosophical ideas, this comparison would be entirely uninteresting, were it not for one thing: the two fields are closely connected. Both sides of the Frankfurt debate agree that an agent choosing C can be responsible for some outcome O only if C is somehow causally connected to O. More specifically, the overwhelming majority of authors agrees that this causal connection consists precisely of actual causation.[1]

On top of this very specific connection between responsibility and causation, the two fields are also more generally related. The literature on Frankfurt cases and their relation to the PAP is flooded with causal language. Moreover, more often than not, the crux of the argument invokes the causal relations that hold between an action performed by the agent and the events which precede it. Therefore it is hard to overestimate the role played by causation in the Frankfurt debate. This makes it all the more striking that for the most part, the advances in the causation literature have been entirely ignored in the literature on Frankfurt cases.[2] It is our aim to set this straight, by applying the developments from the causation literature to the Frankfurt cases and their relation to the PAP.

On our view, the advances in the study of causation are essentially twofold. First, due to the growing interest in causation from the Artificial Intelligence community, causal relations can now be accurately represented by the use of

[1] To our knowledge, Sartorio (2004) is the only exception to this. In a follow up paper we develop a position that also disagrees with this majority, although for different reasons.

[2] Again Sartorio (2016) is a notable exception. However she refrains from using causal models and from giving an explicit definition of causation. This makes her approach quite different from ours.

formal languages. (Structural equations modelling is the most popular language for doing so, however there exist a variety of alternatives (Spirtes et al 2000; Pearl 2000; Vennekens et al 2009).) Second, there has been a substantial improvement in the definitions of causation expressed in such languages. The goal of the current paper is to apply the former advance to the stalemate in the Frankfurt debate. The latter advance will be the subject of a follow-up paper, in which we construct a worked out formal definition of responsibility.

Concretely, in this paper we analyse the Frankfurt debate by making use of causal models. We do so without engaging with the metaphysical positions that often drive the debate (regarding (in)determinism, (in)compatibilism, source vs. leeway, etc.). This allows us to assess the compatibility of the Frankfurt cases and the PAP unhindered by a bias towards any particular metaphysical agenda. Although our analysis results in the position that a moderate version of the PAP is compatible with all of the Frankfurt cases, defending this position is not our primary objective. Rather, we aim to show that the advantage of using causal models lies in the clarity that they bring to an otherwise hopelessly muddled debate.

Specifically, once one has committed to a particular causal model as an appropriate representation of an informal Frankfurt case, it becomes relatively straightforward to judge its relation to various versions of the PAP. As is the case with any exercise in formalisation, there is plenty of room for debate as to whether some causal model is successful at capturing the intended causal structure of the example. However, that debate can be separated entirely from the Frankfurt debate itself, and the causation literature offers plenty of guidance on this matter. Therefore our main goal is simply to argue that the Frankfurt debate cannot be settled without being explicit about the background causal models for the examples under discussion.

Due to the continuous back and forth between opponents and proponents of the PAP, there now exists an abundance of Frankfurt cases. Each case is constructed in order to incorporate objections made to previous cases, which has led to an increasing level of complexity throughout the years. Although we contend that our approach applies to all relevant cases, in the absence of any formal description of what constitutes a Frankfurt case, we are forced to make do with a well-chosen sample of cases. We choose to focus first on a paradigmatic case by Fischer (1999) that has the same structure as the original one by Frankfurt (1969). The discussion over this example is illustrative of the current stalemate, because Fischer first introduced it in 1982, but still reaches back to it almost thirty years later, in order to present what is in essence the same argument (Fischer 1982; 2010).

Due to the sustained and diverse criticism from the defenders of the PAP towards these types of examples, opponents of the PAP have produced more complicated scenarios which are designed with the specific goal of eliminating each and every weakness that simpler cases suffered from. This evolution of evermore complicated cases is currently exemplified by a widely discussed recent case by Pereboom (2009), which makes it an appropriate second test case for our approach.

This paper is structured as follows. In the next section we introduce two versions of the PAP, **Strong PAP** and **Weak PAP**. Section 3 analyses the first of our Frankfurt cases, followed by an analysis of our second Frankfurt case in Section 4. The result will be that **Weak PAP** survives both cases unscathed, whereas the faith of **Strong PAP** depends on the manner in which two different types of responsibility are distinghuished. Further, we propose a third version of the PAP, **Moderate PAP**, and conclude that it can handle both Frankfurt cases, on the condition that one accepts a suggested improvement of Pereboom's definition of robustness. The impact of our analysis is discussed in the Conclusion.

2. Principle of Alternative Possibilities

2.1. Counterfactual Dependence

Before presenting the versions of the PAP that are currently in vogue, we briefly sketch the connection between the two influential articles mentioned above. To do so, we first introduce the notion of *counterfactual dependence*, as it plays a key role in both of them.

Definition 1. *Given that E and C occurred, E is said to be* counterfactually dependent *on C if E had not occurred without C.*

Dating back at least to Hume (1748), it has often been taken for granted that causation can be defined simply as counterfactual dependence. Lewis (1973) convincingly shows this definition to be untenable. He also proposes a definition of causation that is still closely related to counterfactual dependence. Specifically, he presents *Early Preemption* examples to show that counterfactual dependence is not a *necessary condition* for causation, although it is *sufficient*.

The similarity to the article by Frankfurt (1969) is striking. Frankfurt presents a case where all non-causal conditions for an agent being responsible for an act A, whatever they may be, are assumed to be fulfilled. In this initial setup, the act A is counterfactually dependent on some event that is under the agent's control, and it is intuitively clear that the agent is responsible for A. This is in line with the strong intuition that as far as the causal conditions are concerned, counterfactual dependence is sufficient for responsibility. Then Frankfurt modifies the example so that it becomes a case of *Early Preemption*, that is, an example that is structurally identical to the cases discussed by Lewis, to show that counterfactual dependence is not a necessary condition for responsibility.

To be clear, Frankfurt himself does not use the term "counterfactual dependence". Instead, his examples target the PAP, which states that "a person is morally responsible for what he has done only if he could have done otherwise" (Frankfurt 1969: p. 829). Taken literally, this statement mentions only one event, namely that which the agent has done, and therefore the connection with counterfactual dependence may not be immediately obvious. However, Frankfurt distinguishes between an agent performing an act on his own or as the result of coercion. In his examples, the agent has no control over whether or not he performs the act, but he does control whether or not he performs the act on his own or as the result of coercion. This allows us to connect the two by reformulating Frankfurt's version of the PAP as stating that "a person is morally responsible for what he has done only if what he did was counterfactually dependent on something that is under his control".

As far as we know, this version of the PAP has been abandoned even by its most ardent supporters. We suspect that at least indirectly this may have something to do with the universal rejection of the analogous claim that counterfactual dependence is necessary for causation. We now turn to two more nuanced versions of the PAP, which can be seen as two different ways of weakening Frankfurt's version, in an attempt to avoid his criticism.

2.2. **Weak PAP** *and* **Strong PAP**

There are (at least) two different principles about alternative possibilities going around in the current literature. The first states that in order to be *non-derivatively* responsible for some act A, the agent must have been able to prevent it:

Principle 1 (Strong PAP). *If an agent is non-derivatively responsible for performing an act A, then she could have voluntarily not performed A.*

This principle – in various formulations – is the one that has attracted most attention, and is considered the primary target of a Frankfurt-style argument.

Note that the scope of this principle is limited to cases of non-derivative responsibility, which are to be distinguished from cases of *derivative* responsibility. Therefore any Frankfurt case in which the agent is only derivatively responsible for her act is dead on arrival as far as this principle goes. Despite the fact that this makes the distinction a crucial element of the debate, it is often overlooked. The main reason for this oversight is that defenders of **Strong PAP** who invoke this distinction have failed to formulate it in a clear and explicit manner. Widerker, originally a defender of the PAP, introduces it only casually in a footnote, by means of an example (Widerker 2006: p. 163):[3]

[3] Widerker used to be a notable defender of **Strong PAP**, but flipped sides in the paper that

> An agent is *directly* or *non-derivatively* blameworthy for performing an act V only if he is blameworthy for doing so, but not in virtue of being blameworthy for some other act or fact. Otherwise he is *indirectly* or *derivatively* blameworthy for doing V. A typical case of derivative culpability is a scenario in which an agent, who is aware that doing V at T is morally wrong, deliberately places himself in circumstances where he loses his power to avoid doing V at T. If ultimately, the agent does V at T, we say that he is derivatively blameworthy for doing V at T, even though (shortly before T) he could not have avoided doing so.

Without a precise, non-question begging, formulation of the distinction between derivative and non-derivative responsibility, it is impossible to know the scope of **Strong PAP**, and hence we cannot assess whether it holds in general. Nevertheless we can assess if **Strong PAP** is falsified by Frankfurt cases by considering both the possibility that the responsibility in question is non-derivative and that it is derivative. Therefore, our assessment will be conditional on which kind of responsibility we are dealing with.

Of course it is perfectly legitimate for an opponent of **Strong PAP** to doubt that any precise criterion for non-derivativeness can be given, and state that until such doubts have been addressed the principle remains vacuous. We agree that this places the burden of proof firmly on the defenders of **Strong PAP**, but it is important to remark that this line of attack moves the centre of the debate away from Frankfurt cases towards these two notions of responsibility. Our focus here is on the Frankurt cases themselves, so for the purposes of this paper we cast this doubt aside.

The second principle about alternative possibilities is more subtle than the first, stating that a responsible agent should have had *some* alternative that allowed her to avoid being responsible. The subtlety lies in the absence of specifying what the alternative should consist in, except that it should be voluntary.

Principle 2 (Weak PAP). *If an agent is responsible for performing an act A, then she could have* voluntarily *done something such that she would not have been responsible for A.*

Compared to **Strong PAP**, we see that the scope of **Weak PAP** is no longer restricted to cases of non-derivative responsibility. In this sense, naming the principle "Weak" is a misnomer, as increasing the scope of a principle makes it stronger. However if we ignore the scope and compare only the conditions that a case of responsibility has to satisfy, then it is clear that **Weak PAP** is indeed weaker than **Strong PAP**: an agent cannot be responsible for an act that she never

we are quoting here (Widerker 1995; 2000). Palmer (2014) gives the same definition of derivative responsibility in his argument for **Strong PAP**, and likewise introduces it in a footnote only.

performed. Concretely, if the agent could have not performed A – **Strong PAP** – then obviously she could have done something such that she would not have been responsible for A – **Weak PAP**. Therefore anyone who accepts **Strong PAP** also accepts **Weak PAP** for cases of non-derivative responsibility.

Further, whatever one's definition of derivative responsibility, it sounds plausible that if an agent's responsibility for an act A is derived from her responsibility for B, then the former responsibility is avoided whenever the latter is. This implies that most accounts which accept **Strong PAP** will accept **Weak PAP** for derivative responsibility as well, which is why it makes sense to view the second principle as a weaker version of the first.

Some defenders of the PAP only endorse **Weak PAP**, or variants thereof (McKenna 1997; Wyma 1997; Otsuka 1998; Braham & van Hees 2012). Others use the intuitive appeal of **Weak PAP** to justify **Strong PAP**, which suggests that they do not consider there to be much difference between the two (Ginet 1996; Widerker 2000; Palmer 2014). It will turn out, however, that **Strong PAP** is much more vulnerable to Frankfurt cases than **Weak PAP** is. Further on we also introduce a third version of the PAP – **Moderate PAP** – that lies in between the two presented here.

3. The Old School Frankfurt Debate

Here's a paradigmatic Frankfurt case by Fischer (1999: p. 109):

Example 1 (Voting). *Suppose Jones is in a voting booth deliberating about whether to vote for Gore or Bush. (He has left this decision until the end, much as some restaurant patrons wait until the waiter asks before making a final decision about their meal.) After serious reflection, he chooses to vote for Gore and does vote for Gore by marking his ballot in the normal way. Unbeknownst to him, Black, a liberal neurosurgeon working with the Democratic Party, has implanted a device in Jones's brain which monitors Jones's brain activities. If he is about to choose to vote Democratic, the device simply continues monitoring and does not intervene in the process in any way. If, however, Jones is about to choose to vote (say) Republican, the device triggers an intervention which involves electronic stimulation of the brain sufficient to produce a choice to vote for the Democrat (and a subsequent Democratic vote).*

How can the device tell whether Jones is about to choose to vote Republican or Democratic? This is where the "prior sign" comes in. If Jones is about to choose at T_2 to vote for Gore at T_3, he shows some involuntary sign – say a neurological pattern in his brain – at T_1. Detecting this, Black's device does not intervene. But if Jones is about to choose at T_2 to vote for Bush at T_3, he shows an involuntary sign – a different neurological pattern in his brain – at T_1. This brain pattern would trigger Black's device to intervene and cause Jones to choose at T_2 to vote for Gore and to vote for Gore at T_3.

Opponents of the PAP use this example to build up the following argument. If there were no neurosurgeon present, then this example would be entirely unproblematic for defenders of the PAP: Jones is responsible for voting for Gore, and he had the ability to vote for Bush instead. By adding the neurosurgeon, the ability to vote otherwise is removed, leaving Jones with no alternative possibilities. Yet, intuitively, our judgment that Jones is responsible for voting for Gore remains unchanged. Indeed, given that Black's device remained idle the entire time, Jones came to his decision all by himself, just as he would have without Black being present. It seems counterintuitive to suggest that Jones loses responsibility by the mere presence of Black's device. Hence, the argument goes, the ability to do otherwise is not a necessary condition for responsibility.

In turn, defenders of the PAP have argued that Jones in fact *did* have an alternative possibility. If Jones were about to choose to vote for Bush at T_1 instead of Gore, then he would have exhibited a different involuntary sign. The fact that Jones possessed the alternative possibility of exhibiting a different sign, has been coined the "flicker of freedom" (Fischer 1999). It plays a vital role in the debate, as it offers a final line of defense for defenders of the PAP.

In response, opponents of the PAP counter that such a flicker of freedom, as the name suggests, is too flimsy to count as a genuine alternative possibility. For example, Fischer states that "The power involuntarily to exhibit a different sign seems to me to be insufficiently robust to ground our attributions of moral responsibility" (Fischer 1999: p. 110). Although this is a fair point, we want to stress that this issue becomes irrelevant once we get a better picture of the causal model. The debate as we have sketched it here misses out on a central feature of the example, which becomes apparent if we analyse the causal model.

Let $Vote = 1$ denote the event of Jones voting for Gore. A first observation to make is that the neurosurgeon in no way affects the usual mechanism between Jones choosing to vote one way or the other, and Jones effectively voting one way or the other: if Jones chooses to vote for Gore, which we write as $Choice = 1$, then Jones does in fact vote for Gore, and vice versa. In other words, as in any regular example of rational decision-making, the agent performing an act is counterfactually dependent on the agent choosing to perform that act. That is, $Vote = 1$ is counterfactually dependent on $Choice = 1$.

To be clear, this means that Jones's act ($Vote = 1$) is *completely determined* by his choice ($Choice = 1$). This observation in itself already undermines the argument sketched above, for it reveals that Jones's responsibility for $Vote = 1$ is entirely *derived* from his responsibility for $Choice = 1$.[4] Since defenders of the **Strong PAP** admit that its scope is limited to cases of *non-derivative* responsibility, taken literally the argument against the PAP here presented is a straw man.

[4] Even in the absence of a definition of derivative responsibility, we take it to be uncontroversial that this case should fall under it.

Of course one could respond that we are merely splitting hairs, because the focus of the argument can easily be shifted from Jones's responsibility for his actual vote to his responsibility for his choice on how to vote: it is equally intuitive that he remains responsible for $Choice = 1$, despite the absence of alternatives. That may be so, but the fact that this trivial issue is usually overlooked is in itself indicative of the lack of attention for the distinction between derivative and non-derivative responsibility, on which the entire debate hinges. For the sake of argument, let us ignore this problem and follow through on the offered suggestion that the argument should focus on Jones's responsibility for $Choice = 1$.

In order to get a proper understanding of the example, we need to fill in some of the details that are implicit in the causal model that forms the background to our example. Besides Jones's vote ($Vote = 1$) and the choice that he makes ($Choice = 1$), there is one more actual event and one actual omission that are mentioned explicitly: Jones exhibiting some involuntary sign, denoted $Sign = 1$, and Black's device not being triggered, denoted $Device = 0$. Without loss of generality we may assume that the four variables we have introduced are binary, so that to each of the actual events (and omission) there corresponds a counterfactual event: Jones voting for Bush ($Vote = 0$), Jones choosing to vote for Bush ($Choice = 0$), Jones exhibiting a different involuntary sign ($Sign = 0$), and Black's device being triggered ($Device = 1$).

Typically, discussions of this example (and many others like it) assume that these events and the causal relations between them suffice to capture all relevant details of the example. We claim that this is mistaken, for these events leave out the most important part of the causal model, namely the event that gets the entire story going. This becomes apparent once we try to write out the causal model for the variables that we have introduced so far.

It is common practice to model the mechanisms that causally determine a variable X by an equation of the form $X := f(\vec{Y})$, where $f(\vec{Y})$ expresses how the values of certain other variables $\vec{Y}$ determine the value of X.[5] For example, the fact that *Vote* is counterfactually dependent on *Choice* is modelled by taking $Vote := Choice$ as the equation that determines *Vote*. Contrary to the usual equations one encounters in mathematics, there is an asymmetry here between the right and left side of the equation: the expression on the right side determines the value of the variable on the left, but not vice versa. (The use of $:=$ instead of $=$ is intended to highlight this point.) So one should read this equation as stating that changing the value of *Choice* – keeping all else fixed – changes the value of *Vote*, but not vice versa. (We refer the reader to (Pearl 2000) for details on the semantics of these structural equations.)

[5] At least this is common practice if one uses the popular framework of structural equations modelling as presented by Pearl (2000). However our analysis does not depend on using this particular framework.

The equation for *Choice* forms the most crucial part of the causal model. Let us start with considering the causal model for the normal setting, i.e., the setting in which Black the neurosurgeon and his device are absent. The example states that in the given context, Jones's choice (*Choice*) is perfectly correlated with him exhibiting a particular involuntary sign: if $Sign = 1$, we have $Choice = 1$, and if $Sign = 0$, we have $Choice = 0$. Whenever two events are perfectly correlated, we either have that one is a cause of the other, or they both have a common cause. Obviously Jones's choice does not cause him to exhibit the sign, since the sign is exhibited at T_1 and the choice occurs at T_2. Neither are we to imagine that the involuntary sign is a cause of his choice. This is betrayed by the meaning of the term "sign" and confirmed by the stipulation that the sign is exhibited involuntary, implying that it is merely a side-effect of Jones's deliberation.[6] Therefore there must be some event preceding $Sign = 1$ that is the cause of both Jones's exhibiting the sign and of him choosing to vote for Gore.[7] Let us call this event $Reflection = 1$, in line with the statement that Jones makes his choice "After serious reflection".

So what are the causal relations between these three events? Given that *Sign* and *Choice* are perfectly correlated, and that their causal connection is mediated entirely through $Reflection$, we conclude that both $Sign = 1$ and $Choice = 1$ must be counterfactually dependent on $Reflection = 1$. In other words, the equations for *Sign* and *Choice* are simply $Sign := Reflection$ and $Choice := Reflection$. These equations capture both the fact that Jones reflecting as he actually did resulted in him choosing to vote for Gore and to exhibit a particular involuntary sign, as well as the counterfactual that if Jones had reflected differently ($Reflection = 0$) he would have chosen to vote for Bush and he would have exhibited a different involuntary sign. All we have done by introducing these variables and equations is to make explicit certain details of the example that are hidden in plain sight.

To summarize, an appropriate causal model for our example in the absence of Black and his device is given by the following three equations:

$$\begin{aligned} Vote &:= Choice. \\ Choice &:= Reflection. \\ Sign &:= Reflection. \end{aligned}$$

Since this model consists entirely of relations of counterfactual dependence, we refer to it as *Dependence*. Before moving on to describe the causal model for the full example, let us pause to analyse this model. By assumption, Jones had an

[6] We point out that our analysis in no way depends on the assumption that the sign is not itself a cause of Jones's choice. A causal model in which the equation for choice is $Choice := Sign$ works just as well as the more intuitive one that we present here.

[7] In principle there might be several events that together form a cause rather than just a single one, but we can ignore this complication.

alternative possibility for $Vote = 1$ in the simplified story under consideration. Given that *Vote* is determined entirely by *Choice*, which is in turn determined entirely by $Reflection$, the alternative possiblity for $Vote = 1$ must be entirely dependent on Jones having the alternative possibility of having reflected differently ($Reflection = 0$). Therefore Jones had direct voluntary control over $Reflection$, which indirectly gave him voluntary control over *Choice* (and over *Vote*).

We deliberately take no position on what it means for an agent to have "direct voluntary control", because this is irrelevant for the current analysis. Given that Jones's relation to $Reflection$ is in no way influenced by the addition of Black's device, the Frankfurt argument has no bearing whatsoever on whether or not Jones had direct voluntary control over $Reflection$. This fact is obscured if one focusses exclusively on the sign – which is involuntary – and ignores its cause, as Fischer does.

Given the ambiguity in the literature on what constitutes derivative responsibility, we want to keep an open mind and consider it an option that Jones's responsibility for $Choice = 1$ is derivative. (Recall that the scope of **Strong PAP** is limited to cases of non-derivative responsibility.) If Jones's responsibility for $Choice = 1$ is derivative of his responsibility for $Reflection = 1$, then **Weak PAP** is confirmed in this example for $Choice = 1$ and **Strong PAP** is confirmed for $Reflection = 1$. If, on the other hand, Jones's responsibility for $Choice = 1$ is non-derivative, then **Strong PAP** is confirmed for $Choice = 1$.

Now let us consider what happens to the causal model when we add Black and his device to the story. First there is the equation for *Device*, which is again very simple: the device being triggered is counterfactually dependent on Jones exhibiting a sign that is different from the particular sign that he exhibited in the actual story, so we get $Device := \neg Sign$. (We make use of the standard propositional connectives: $\neg$ for negation, $\vee$ for disjunction, and $\wedge$ for conjunction.) Second, note that the causal mechanisms for *Vote* and *Sign* are unaffected. Only Jones's choice is influenced by the device, in the following manner. If the device is triggered, then it overrules the normal functioning of Jones's deliberation by ensuring that Jones chooses to vote for Gore. If the device remains idle, as is the case in the actual scenario, then the causal mechanism for Jones's choice behaves as usual. This is captured by the equation $Choice := Device \vee Reflection$.

To sum up, the following is an appropriate causal model for the full example:

$$\begin{aligned} Vote &:= Choice. \\ Sign &:= Reflection. \\ Choice &:= Device \vee Reflection. \\ Device &:= \neg Sign. \end{aligned}$$

To those familiar with the literature on causation, the relation between $Reflection = 1$ and $Choice = 1$ in this model should be easily recognised as a case of *Early Pre-*

emption. Such cases have been discussed ad nauseam, because they form the most famous counterexample to the necessity of counterfactual dependence for causation: although $Choice = 1$ is not counterfactually dependent on $Reflection = 1$, there is widespread agreement that $Reflection = 1$ is an actual cause of $Choice = 1$ nonetheless.[8] Therefore in causal terms a Frankfurt case can be characterised by stating that we move from *Dependence* to *Early Preemption*.

As noted, in the absence of Black's device, Jones possessed an alternative possibility for $Choice = 1$ in virtue of him possessing an alternative possibility for $Reflection = 1$. How does the addition of Black's device change this? Jones still had direct control over $Reflection$, and therefore it remains the case that he could have reflected differently ($Reflection = 1$). Had he done so, however, the device would have been triggered, and Jones would have chosen to vote for Gore all the same. In short, Jones loses control over *Choice*, while retaining his control over $Reflection$.

What conclusions can we draw from this? First, we consider **Strong PAP**. Jones had an alternative possibility for $Reflection = 1$, but not for $Choice = 1$. This means that if Jones's responsibility for $Choice = 1$ is non-derivative, then the Frankfurt argument is successful, insofar as **Strong PAP** is its intended target. However, if his responsibility for $Choice = 1$ is derived from Jones's responsibility for $Reflection = 1$, then it falls beyond the scope of **Strong PAP** and the argument fails to falsify it.

Second, we consider **Weak PAP**. In this case, the addition of Black's device doesn't change a thing, and the Frankfurt argument falls apart. How so? **Strong PAP** remains confirmed for $Reflection = 1$, because it remains true that it was open to Jones to voluntarily reflect differently ($Reflection = 0$). If he had, then he would have been*forced* by the device to choose to vote for Gore, rather than make that choice voluntarily. Despite the many different views on responsibility, there is a widespread consensus that an agent cannot be responsible for something that she was forced to do. Therefore Jones would not have been responsible for voting for Gore in this counterfactual scenario. In other words, there existed an alternative possibility such that he voluntarily could have avoided being responsible for $Choice = 1$, confirming **Weak PAP**.

As a result of our analysis, we get a more nuanced picture of the Frankfurt argument, which offers the following lessons. First, the force of the argument with regard to **Strong PAP** cannot be assessed without explicit (non-questionbegging) criteria to distinguish between derivative and non-derivative responsibility. Second, if it turns out that $Choice = 1$ is a case of non-derivative responsibility, then **Strong PAP** is falsified, but **Weak PAP** is not. Third, if it turns out that $Choice = 1$ is a case of derivative responsibility, then neither **Strong PAP** nor **Weak PAP** are

[8] Strictly speaking, our own preferred definition of causation forms an exception, however that proviso is of no concern for our current purposes (Beckers & Vennekens 2017a;b).

falsified by the Frankfurt argument.

Perhaps some readers are not convinced by our analysis of *Voting*, because they adhere to a metaphysical position regarding causal determinism and its relation to free will that precludes the causal models here suggested. Although we have tried to avoid committing ourselves to any particular metaphysical position, we do not wish to rule out this possibility. Furthermore, it is an intrinsic feature of representing causal relations by means of causal models that there is room for disagreement about whether some model is an appropriate formalisation of a natural language description (Halpern & Hitchcock 2010). The point of using causal models is not to rule out discussion regarding the causal relations, but rather to provide the tools that allow this discussion to proceed in a more perspicuous fashion. We welcome scepticism regarding these particular causal models, on the condition that it is accompanied by the acknowledgement that a debate on Frankfurt cases cannot move forward without being explicit about the causal models involved.

To amplify this last point, we present an example by Mele and Robb (1998) that Fischer cites in support of his own view. According to him, this is yet another example that successfully undermines the PAP. As it turns out, however, different elements of the example suggest two very different causal models (Fischer 1999: p. 115).[9]

Example 2 (Stealing). *At T_1, Black initiates a certain deterministic process P in Bob's brain with the intention of thereby causing Bob to decide at T_2 (an hour later, say) to steal Ann's car. The process, which is screened off from Bob's consciousness, will deterministically culminate in Bob's deciding at T_2 to steal Ann's car unless he decides on his own at T_2 to steal it or is incapable at T_2 of making a decision (because, e.g., he is dead by T_2). (Black is unaware that it is open to Bob to decide on his own at T_2 to steal the car; he is confident that P will cause Bob to decide as he wants Bob to decide.) The process is in no way sensitive to any "sign" of what Bob will decide. As it happens, at T_2 Bob decides on his own to steal the car, on the basis of his own indeterministic deliberation about whether to steal it, and his decision has no deterministic cause. But if he had not just then decided on his own to steal it, P would have deterministically issued, at T_2, in his deciding to steal it. Rest assured that P in no way influences the indeterministic decision-making process that actually issues in Bob's decision.*

In a nutshell, the idea here is that the indeterminism of the process that actually results in Bob's decision ensures that Bob would have had access to alternative possibilities were it not for Black, whereas the determinism of the process P ensures that *all* alternative possibilities are removed (including any flicker of freedom). Still, the argument goes, the presence of the process P does

[9] A similar point can be made about another such example that Fischer discusses, namely the one by Stump (1999).

not affect our intuitive judgment that Bob is responsible for stealing Ann's car, showing that alternative possibilites are not necessary for responsibility.

What are we to make of this example? As before, we introduce binary variables to model all relevant events. Let $Steal = 1$ stand for Bob deciding to steal Ann's car. We take $Self = 1$ to represent Bob's deliberation about whether to steal the car. The stipulation that this event is "indeterministic", is meant to capture the assumption that Bob had the ability to deliberate otherwise, denoted by $Self = 0$. So in normal circumstances, if Black had not been present, Bob's decision to steal is counterfactually dependent on his deliberation process: $Steal := Self$.

Now we add Black. Let $Init = 1$ stand for Black initiating the deterministic process P at T_1, while $Culm = 1$ stands for the process P culminating in Bob deciding at T_2 to steal the car. Both the indeterministic process and the deterministic process are by themselves sufficient in determining Bob's decision to steal the car, so the equation for *Steal* becomes: $Steal := Self \lor Culm$.

However, there is an asymmetry between both processes, since it is stated that "The process [P], which is screened off from Bob's consciousness, will deterministically culminate in Bob's deciding at T_2 to steal Ann's car *unless he decides on his own at T_2 to steal it* [emphasis added]". In other words, the event $Culm = 1$ can occur only if Bob is not about to decide by himself to steal the car ($Self = 1$). This suggests the following equation for the culmination of P: $Culm := Init \land \neg Self$.

Taken together, this gives the following causal model:

$$Steal := Self \lor Culm.$$
$$Culm := Init \land \neg Self.$$

The relation between $Self = 1$ and $Steal = 1$ in this model is almost structurally isomorphic to the relation between $Reflection = 1$ and $Choice = 1$ that we encountered in our model for *Voting*. The only differences between the two models are the presence of the intermediate event $Sign = 1$ in the latter, and the addition of *Init*. These differences are irrelevant for categorising the example as a case of *Early Preemption*. Therefore, this example is identical to *Voting*, and our earlier analysis can be repeated.

But it appears that we have ignored a key feature of the example! The deterministic process P is presumed to be entirely independent of the indeterministic process of Bob's deliberation: "The process [P] is in no way sensitive to any "sign" of what Bob will decide". This statement suggests that once the process P has been initiated, it culminates in forcing Bob to decide to steal no matter what, in which case we would have the following causal model:

$$Steal := Self \lor Culm.$$
$$Culm := Init.$$

This model is associated with another paradigmatic case study from the causation literature, namely *Symmetric Overdetermination*.

It has been well-established that cases of *Early Preemption* are very different from cases of *Symmetric Overdetermination*, and therefore these two models are in clear opposition to one another.[10] As before, we do not wish to rule out the possibility that a more sophisticated model can be given which better captures the example at hand. However, until such a model has been suggested, this example stands as an illustration of the ambiguity that is prevalent in the literature on Frankfurt cases. A debate that crucially depends on causal relations cannot move forward if people disagree about the causal relations to begin with.

4. The Current Frankfurt Debate

One might object that our analysis is too simple, because by now there are far more complex Frankfurt cases than *Voting*. To meet this objection we have a look at one of the most recent and complex Frankfurt cases, which was formulated by Pereboom (2009) as an attempt to counter arguments that were made against simpler Frankfurt cases. His example builds upon *Voting*, but it adds a subtle nuance which makes the alternative possibility less straightforward. He offers several versions of the example, but for the purposes of this paper the differences between them can be ignored (Pereboom 2009: p. 113).

Example 3 (Tax Evasion)**.** *Joe is considering claiming a tax deduction for the registration fee that he paid when he bought a house. He knows that claiming this deduction is illegal, but that he probably won't be caught, and that if he were, he could convincingly plead ignorance. Suppose he has a strong but not always overriding desire to advance his self-interest regardless of its cost to others and even if it involves illegal activity. In addition, the only way that in this situation he could fail to choose to evade taxes is for moral reasons, of which he is aware. He could not, for example, choose to evade taxes for no reason or simply on a whim. Moreover, it is causally necessary for his failing to choose to evade taxes in this situation that he attain a certain level of attentiveness to moral reasons. Joe can secure this level of attentiveness voluntarily. However, his attaining this level of attentiveness is not causally sufficient for his failing to choose to evade taxes. If he were to attain this level of attentiveness, he could, exercising his libertarian free will, either choose to evade taxes or refrain from so choosing (without the intervener's device in place). However, to ensure that he will choose to evade taxes, a neuroscientist has, unbeknownst to Joe, implanted a device in his brain, which, were it to sense the requisite level of attentiveness, would electronically stimulate the right neural centers so as to inevitably result in his making this choice. As it happens, Joe does not*

[10] Essentially the same criticism is given by Widerker (2000), however he does not formulate it in this manner, nor does he use causal models.

attain this level of attentiveness to his moral reasons, and he chooses to evade taxes on his own, while the device remains idle.

Intuitively, Joe is responsible for choosing to evade taxes ($Choice = 1$). Yet it was clearly impossible for him to choose not to evade taxes ($Choice = 0$). So as before, if his responsibility for $Choice = 1$ is non-derivative, then this example falsifies **Strong PAP**. If not, then it falls beyond the scope of **Strong PAP**.

The only alternative that was open to him was to voluntarily become attentive to moral reasons ($Attention = 1$). If he had done so, then the device would have kicked in ($Device = 1$), which would have forced Joe to choose to evade taxes. Since he would have been forced, under this alternative he would end up not being responsible for choosing to evade taxes. Therefore as with *Voting*, this example confirms **Weak PAP**. In both these respects *Tax Evasion* adds nothing new to our earlier analysis.

Indeed, as with *Voting*, the full scenario can be appropriately modelled as a standard case of *Early Preemption*:

$$Choice := \neg Attention \vee Device.$$
$$Device := Attention.$$

Pereboom agrees that his example does not challenge **Weak PAP**, but claims that the mere availability of a voluntary alternative fails to meet the bar for a sensible version of the PAP. On his view, any interesting version of the PAP requires there to be a properly *robust* alternative. Therefore the focus of his attack is a version of the PAP that lies somewhere in between **Weak** and **Strong PAP**, which we can characterise as follows:

Principle 3 (Moderate PAP). *If an agent is responsible for performing an act A, then she could have done something* robust *such that she would not have been responsible for A.*

Since the only difference with **Weak PAP** is the requirement that the alternative be robust rather than merely voluntary, the success of Pereboom's example depends entirely on how we should define robustness. Pereboom argues for accepting the following informal definition, which he constructed in order to meet the objections raised against earlier versions that were formulated by several proponents of the PAP (Pereboom 2009: p. 112):

Definition 2 (Robustness (1)). *For an alternative possibility to be relevant per se to explaining why an agent is morally responsible for an action it must satisfy the following characterization: she could have willed something different from what she actually willed such that she has some degree of cognitive sensitivity to the fact that by willing it she thereby would be, or at least would likely to be, precluded from the responsibility she actually has.*

This definition highlights an aspect of the discussion that we have ignored thus far, namely the epistemic state of the agent. Pereboom – as do most others – assumes that one key feature of being responsible consists in being aware of certain relevant facts regarding the available possibilities. It would clearly be too much to demand that the agent has an accurate grasp of the entire situation, for we cannot expect her to know each and every detail of the causal model. The idea behind robustness is that in the least, the agent should have been aware of the fact that she voluntarily refused to choose an alternative that might have precluded her from being responsible. As Pereboom points out, this idea is commonly used to motivate the PAP (Ginet 1996; Otsuka 1998; Moya 2006).

Note, however, that according to Definition 2, the mere existence (and recognition) of an alternative which *might* avoid responsibility does not suffice for robustness. Rather, the alternative has to be such that the agent "would *likely to be* precluded from the responsibility she actually has". Pereboom justifies this stricter condition by pointing out that a healthy dosage of scepticism would almost always ensure there to be some alternative which has a non-zero probability of occurring, and yet such unlikely alternatives hardly come to mind when we are considering responsibility. On the other hand, he claims, certainty regarding the alternative also seems like too much to ask, since a probability of 0.95, for instance, sounds like more than enough. However "the threshold probability, as one would expect, is difficult or impossible to determine", which is why he settles for the vague characterisation of the alternative *being likely* (Pereboom 2009: p. 112). We claim that a better solution presents itself once we make use of causal models.

First let us consider a simplified version of the normal, pre-intervention, *Tax Evasion* scenario that is analogous to *Voting*. We forget about the complication that even if Joe were to become attentive to moral reasons, he might still choose to evade taxes. Rather, we assume that Joe's choice is counterfactually dependent on him not becoming attentive to moral reasons, giving the following simple one-equation causal model: $Choice := \neg Attention$. As before, we call this scenario *Dependence*.

Second we consider the opposite scenario, in which Joe would have certainly chosen to evade taxes: were he to become attentive to moral reasons, his self-interest would certainly override these reasons and compel him to choose to evade taxes. Just as the full setup of the actual scenario, this scenario is a case of *Early Preemption*:

$$\begin{aligned} Choice &:= \neg Attention \vee Override. \\ Override &:= Attention. \end{aligned}$$

Finally we consider the actual scenario without the intervening neuroscientist, say *Pre-Tax Evasion*, which lies in between *Dependence* and *Early Preemption*:

$$Choice := \neg Attention \vee Override.$$
$$Override := Attention \wedge Self.$$

Here the binary variable $Self$ represents whether or not Joe would voluntarily decide to let his self-interest prevail over the moral reasons. If Joe believes that $Self$ is certain to be true – $P(Self = 1) = 1$ – then *Pre-Tax Evasion* reduces to *Early Preemption*, and if he believes that $Self$ is certain to be false – $P(Self = 1) = 0$ – then it reduces to *Dependence*. The former scenario is not suitable to build up a Frankfurt case, because Joe wouldn't even have a robust alternative possibility to begin with (i.e., he wouldn't have a robust alternative possibility even without the neuroscientist present). The latter scenario is not suitable, because Joe would still have a robust alternative possibility even with the neuroscientist present (as is the case with *Voting*). Therefore the fact that Joe believes the value of $Self$ to be undetermined – $0 < P(Self = 1) < 1$ – is a crucial feature of Pereboom's example.

However, mere uncertainty regarding $Self$ is not enough for Pereboom's Frankfurt argument to work. To see why, consider the option that Joe believes $Self$ is likely to be false – $P(Self = 1) = low$. In this case, Joe did have a robust alternative with respect to $Attention = 0$ in *Tax Evasion*: Joe correctly believes that if he wills $Attention = 1$, then it is likely that he will be precluded from being responsible of choosing to evade taxes. Of course Joe is mistaken as to why his belief is correct, because he bases it on the false belief that if he wills $Attention = 1$, then he will most likely not choose to evade taxes at all. But unless we are to make the patently question-begging assumption that an agent needs to be aware of the presence of an undetectable device in order for her to have a robust alternative, the source of Joe's belief is entirely irrelevant.

Therefore, a further crucial aspect of Pereboom's approach is that he defines an alternative to be robust only if it is likely, as opposed to merely possible. Concretely, it is crucial for his approach that Definition 2 distinguishes between $P(Self = 1) = high$ and $P(Self = 1) = low$. As mentioned, he defends this choice by stating that it's impossible to find the cut-off point at which the probability of the alternative becomes high enough to qualify as robust.

So far our focus has been exclusively on the counterfactual story, in which we consider what would happen if $Attention = 1$. Specifically, our three scenarios differ only with respect to the probability that the backup mechanism ($Self = 1$) is effective, whereas they all agree on what happens if $Attention = 0$. Given that we are trying to assess the agent's epistemic state regarding an alternative, it is natural to ask: "An alternative to *what*?". To that end, we also add uncertainty to the story in which $Attention = 0$, making the picture fully symmetric. Concretely, we add the binary variable X that represents whether or not $Attention = 0$ is sufficient for $Choice = 1$:

$$Choice := (\neg Attention \wedge X) \vee Override.$$
$$Override := Attention \wedge Self.$$

The agent can voluntarily will either to become attentive to moral reasons, or not. The question that concerns us, is which of these she ought to will in order for her to be "off the hook" (Pereboom 2009: p. 114). If we limit ourselves to scenarios such that the agent believes $P(X = 1) = 1$, as we have been doing, then we are faced with Pereboom's vagueness surrounding a proper cut-off probability for $P(Self = 1)$. However, since we are comparing two alternatives, the obvious solution is to compare the respective probabilities that they result in a particular outcome. In other words, the obvious cut-off point for the alternative to $Attention = 0$ to be robust is when $P(Self = 1) < P(X = 1)$. Elzein (2013) proposes the same solution, stating that "Intuitively, ..., alternatives seem important so long as they are comparatively likely to result in better outcomes. We are much less interested in the question of whether they are likely to result in better outcomes full stop."

Were we to apply this suggestion to *Tax Evasion*, in which it is assumed that $P(X = 1) = 1$, then we would get the result that Joe had a robust alternative possibility after all, regardless of whether $P(Self = 1) = high$ or $P(Self = 1) = low$. The upshot would be that Pereboom's Frankfurt argument is no longer successful, as *Tax Evasion* would fail to falsify **Moderate PAP**.

The only argument Pereboom gives for defining robustness in a manner that excludes unlikely alternatives is that it would set the bar for robustness far too low. Considering an example in which Joe is completely unaware of the fact that him drinking coffee – which is poisonous – would result in him not evading taxes, he states (Pereboom 2009: p. 112):

> Joe might well agree that the probability of this connection [between drinking coffee and not evading taxes] is non-zero – he might admit, for instance, that it's at least .000001, and if he's taken a class in epistemology or probability, something like this might well be his response. But, intuitively, this is not sufficient to generate robustness.

The solution that we have sketched above takes care of this concern. By the same epistemological considerations, Joe would just as well agree that the probability of the connection between not drinking coffee and not evading taxes is also .000001, so that both alternatives are on a par. Therefore our proposed modification to the definition of robustness is equally capable of excluding such undesirable examples as Pereboom's proposal is.

Further, we claim that our proposal is backed up by intuition, as the following example illustrates.

Example 4 (Trolley)**.** *As has become quite common these days, there is a runaway trolley approaching a split in the railroad tracks. Joe is standing next to the split, with his hands on a switch. If he flips the switch, the trolley will be diverted to the left track, which leads to Jones, who is tied to the track. If Joe does not flip the switch, the trolley will continue onto the right track, which leads to another split further down. The left track of this second split forms a loop with the left track of the first split, whereas the right track of the second split goes into another direction entirely. Joe does not know which track the trolley would take at the second split, he believes it might go either way. Joe decides to flip the switch, so that the trolley goes down the left track and kills Jones.*

In this example everyone would agree that Joe made the wrong choice, and the reason for this is precisely that he had a robust alternative possibility: if he had not flipped the switch and the trolley had gone down the right track at the second split, then Jones would not have died. Crucially, this judgment remains true even if we specify that there's only a small probability that the trolley would have gone on the right track on the second split. What matters here, is whether or not the probability of avoiding Jones's death when not flipping the switch is larger than avoiding his death when flipping the switch.[11] As long as that is the case, the only reason Joe could have had for flipping the switch is that he wants Jones to die, which is what explains why we hold him responsible for Jones's death.

This might sound counterintuitive in case the probability of Jones's death is extremely small. However, we're making the idealised assumption here that there are no other outcomes involved whatsoever. For example, it is not the case that flipping the switch is somehow beneficial to Joe in its own right, or that he enjoys making random decisions. Therefore the relative probabilities of the different alternatives are the only determinants of his decision. Obviously things are less straightforward once we drop this assumption, but one should keep in mind that Frankfurt cases revolve around responsibility for a single outcome.

Therefore we suggest the following definition of robustness as an improvement over Definition 2:

Definition 3 (Robustness (2)). *For an alternative possibility to be relevant per se to explaining why an agent is morally responsible for an action it must satisfy the following characterization: she could have willed something different from what she actually willed*

[11] Elzein (2013) gives a very similar example, and she likewise concludes that "all that really seems to matter is that each agent reduces the odds of wrongdoing as much as is reasonably *possible for that agent*, irrespective of where the "likelihood" border lies, or how close to it the alternative gets us. This highlights something important for assessing an agent's blame. We want to know whether the agent did her *reasonable best* to avoid the blameworthy behaviour. "Doing your reasonable best" is not a matter of making a good outcome likely; it's a matter of making the *best possible* outcome *as* likely as you reasonably *can*, given the range of options available to you. This is essentially a comparative rather than absolute matter."

such that she has some degree of cognitive sensitivity to the fact that by willing it she thereby would have been more likely to be *precluded from the responsibility she actually has.*

Since Joe considers it more likely that not attending to moral reasons will result in him being responsible for choosing to evade taxes, than attending to moral reasons, he has a robust alternative in *Tax Evasion*, thus confirming **Moderate PAP**.

4.1. The Timing Objection

Ginet (2002) has proposed a different argument against Pereboom's analysis of *Tax Evasion*, that has been coined "the timing objection". The timing objection is based on the distinction between an agent being responsible for some act *A at time T* and the agent being responsible for *A simpliciter*. Proponents of the timing objection claim that Joe may well be responsible for $Choice = 1$ at some particular time T, without being responsible for $Choice = 1$ simpliciter. On this view, it suffices for a defender of the PAP to show that Joe did have a robust alternative possibility for $Choice = 1$ at time T, as opposed to having to argue for the much stronger claim that Joe had a robust alternative for $Choice = 1$ simpliciter. The success of this approach depends on the plausibility of the assumption that our intuitive judgment regarding Joe's responsibility is in fact an intuition about time-indexed responsibility, rather than responsibility simpliciter. Again we can clarify this debate by invoking the causation literature, because an entirely analogous debate has taken place regarding causation.

Defenders of the timing objection accept that Joe could not have made a different choice ($Choice = 0$), but point out that Joe could have made his choice at a slightly different time T'. This brings us right back to our starting point, namely the claim that the PAP should be interpreted as the claim that counterfactual dependence is necessary for responsibility: the occurrence of $Choice = 1$ at T is counterfactually dependent on something that is under Joe's control, confirming the PAP and averting the Frankfurt threat.

This strategy is reminiscent of a corresponding attempt to salvage the necessity of counterfactual dependence for causation. The underlying assumption in both cases is that our intuitive judgments regarding an event – be they judgments about responsibility or about causation – are sensitive to the precise circumstances, temporally and otherwise, under which the event occurs. Lewis (1986) was the first to suggest that, in some cases, the proper target of our causal judgments is indeed such a "fragile" event. However, any account of causation that takes such fragility to be a property of *all* causal judgments faces an insurmountable challenge, for it leads to an explosion of causal relations: any event which had only the slightest of impacts on the manner in which some outcome occurred, would be counted as a cause. This challenge carries over to responsibility, as the following example illustrates.

Example 5 (Late Preemption)**.** *Suzy and Billy both throw a rock at a bottle. Suzy's rock gets there first, shattering the bottle. However Billy's throw was also accurate, and would have shattered the bottle had it not been preempted by Suzy's throw.*

Late Preemption is another paradigmatic example that has been amply discussed in the causation literature. Our intuitive verdict is pretty clear: Suzy's throw is a cause of the bottle shattering, whereas Billy's throw is not. Since the shattering of the bottle is not counterfactually dependent on Suzy's throw, we have another counterexample to the necessity of counterfactual dependence for causation. As with the timing objection, one could try to avoid this conclusion, by distinguishing between the bottle shattering simpliciter, and it shattering at some particular time *T*. Paul and Hall (2013) provide a comprehensive discussion of this proposal, pointing out the severe problems that it faces. We here present the most obvious one, namely the explosion of causes (Paul & Hall 2013)[p. 106].

> Consider that, e.g., thanks to the gravitational attraction between Billy's rock and Suzy's rock, Billy's throw will make some very slight difference to the time and manner of the bottle-shattering. But this does not qualify it as a cause of the event of the shattering (although it is, arguably, a cause of certain of the shattering's properties). But if, for instance, we took the bottle-shattering to be extremely modally fragile – so that any counterfactual difference in the manner or time of shattering would necessarily yield a numerically different event – then we get the unwelcome result that the shattering depends on Billy's throw, and hence is (according to counterfactual analyses) an effect of that throw.

This criticism transfers to the timing objection for responsibility. To spice up the moral character of the example, assume that the bottle contains some irreplaceable life-saving medicine for the terminally ill Jones. Assuming that both Suzy and Billy fulfill all non-causal conditions for responsibility, whatever they may be, the only possible difference between them has to consist in the causal relations between their respective throws and the shattering of the bottle. Per the timing objection, we ought to consider the shattering of the bottle at slightly different moments as numerically different events. But then we get the extremely counterintuitive conclusion that Billy is also responsible for the shattering of the bottle (as well as the ensuing death of the unfortunate Jones).[12] Therefore

[12] Of course Billy may well be responsible for *intending* to shatter the bottle, and this in itself is enough to consider his action immoral. But that doesn't make him responsible for the outcome itself.

the timing objection regarding responsibility is just as unappealing as its causal counterpart.

Pereboom (2012) has also responded to the timing objection, but in quite a different manner. In turn, Palmer (2013) argues that Pereboom's response does not succeed in undermining the force of the timing objection. There is an important reason why Pereboom cannot dismiss the timing objection in the manner outlined above: he invokes a time-indexed causal judgment himself, in order to fend off the criticism that Joe's responsibility for $Choice = 1$ is derivative and therefore not susceptible to **Strong PAP**. His worry is that $Choice = 1$ could be considered derivative if it was causally determined by Joe not being attentive to moral reasons. To deal with this worry, he focusses on the observation that Joe not being attentive to moral reasons at any particular time T does not causally determine his choice, and ignores the observation that Joe not being attentive to moral reasons simpliciter does causally determine his choice. In this manner his position almost invites the timing objection. In contrast, we have chosen to represent the absence of moral reasons as the single omission $Attention = 0$, in line with the standard practice in the causation literature.[13]

5. Conclusion

In this paper we have illustrated the importance of using causal models to settle the debate over the PAP. Although disagreements over the appropriateness of particular causal models present challenges of their own, these challenges are in no way unique to this debate, and can be separated from the discussion regarding the PAP itself. We analysed three different versions of the PAP and their relation to two paradigmatic Frankfurt cases, concluding that:

- **Weak PAP** is not invalidated;
- **Moderate PAP** is consistent with both cases if one accepts a suggested improvement to a well-known definition of robustness;
- **Strong PAP** cannot be assessed without offering a criterion for distinguishing between two types of responsibility.

In a follow-up paper we aim to provide a positive case for **Moderate PAP**, and build a formal definition of moral responsibility based on it. To do so, we will draw on promising recent work on actual causation.

[13] The temporal properties of omissions are notoriously hard to get right, and lie at the heart of several complicated examples from the causation literaure. We refer the reader to our work on causation for a discussion of this matter (Beckers & Vennekens 2017a).

Acknowledgements

Sander Beckers was funded by a fellowship from the Belgian American Educational Foundation.

References

Beckers S, Vennekens J (2017a) A principled approach to defining actual causation. Synthese forthcoming

Beckers S, Vennekens J (2017b) The transitivity and asymmetry of actual causation. Ergo 4(1):1–27

Braham M, van Hees M (2012) An anatomy of moral responsibility. Mind 121(483):601–634

Elzein N (2013) Pereboom's frankfurt case and derivative culpability. Philosophical Studies 166:553–573

Fischer JM (1982) Responsibility and control. Journal of Philosophy 79:24–40

Fischer JM (1994) The Metaphysics of Free Will. Blackwell

Fischer JM (1999) Recent work on moral responsibility. Ethics 110(1):93–139

Fischer JM (2010) The frankfurt cases: the moral of the stories. The Philosophical review 119(3):315–336

Frankfurt HG (1969) Alternate possibilities and moral responsibility. Journal of Philosophy 66(23):829

Ginet C (1996) In defense of the principle of alternative possibilities: Why i don't find frankfurt's argument convincing. Philosophical Perspectives 10:403–417

Ginet C (2002) Review of pereboom's "living without free will". Journal of Ethics 6:305–309

Halpern J (2016) Actual Causality. MIT Press

Halpern J, Hitchcock C (2010) Actual causation and the art of modeling. In: Causality, Probability, and Heuristics: A Tribute to Judea Pearl, London: College Publications, pp 383–406

Hume D (1748) An Enquiry concerning Human Understanding

Lewis D (1973) Causation. Journal of Philosophy 70:113–126

Lewis D (1986) Causation. In: Philosophical Papers II, Oxford University Press, pp 159–213

Lewis D (2000) Causation as influence. Journal of Philosophy 97(4):182–197

McKenna M (1997) Alternative possibilities and the failure of the counterexample strategy. Journal of Social Philosophy 28:71–85

Mele A, Robb D (1998) Rescuing frankfurt-style cases. Philosophical Review 107:97–112

Moya C (2006) Moral Responsibility: The Ways of Skepticism. Oxford University Press

Otsuka M (1998) Incompatibalism and the avoidability of blame. Ethics 108:685–701

Palmer D (2013) The timing objection to the frankfurt cases. Erkenntnis 78:1011–1023

Palmer D (2014) Deterministic frankfurt cases. Synthese 191(16):3847–3864

Paul L, Hall N (2013) Causation: a user's guide. Oxford University Press

Pearl J (2000) Causality: Models, Reasoning, and Inference. Cambridge University Press

Pereboom D (2009) Further thoughts about a frankfurt-style argument. Philosophical Explorations 12(2):109–118

Pereboom D (2012) Frankfurt examples, derivative responsibility, and the timing objection. Philosophical Issues 22:298–315

Sartorio C (2004) How to be responsible for something without causing it. Philosophical Perspectives 18:315–336

Sartorio C (2016) Causation and Free Will. Oxford University Press

Spirtes P, Glymour C, Scheines R (2000) Causation, Prediction, and Search, 2nd edn. MIT Press

Stump E (1999) Dust, determinism, and frankfurt. Faith and Philosophy 16(3):413–422

Timpe K (2006) The dialectic role of the flickers of freedom. Philosophical Studies 131(2):337–368

Vennekens J, Denecker M, Bruynooghe M (2009) CP-logic: A language of probabilistic causal laws and its relation to logic programming. Theory and Practice of Logic Programming 9:245–308

Weslake B (2015) A partial theory of actual causation. The British Journal for the Philosophy of Science forthcoming

Widerker D (1995) Libertarianism and frankfurt's attack on the principle of alternative possibilities. The Philosophical review 104(2):247–261

Widerker D (2000) Frankfurt's attack on the principle of alternative possibilities: A further look. Philosophical Perspectives 14:181–201

Widerker D (2006) Libertarianism and the philosophical significance of frankfurt scenarios. Journal of Philosophy 103(4):163–187

Woodward J (2003) Making Things Happen: A Theory of Causal Explanation. Oxford University Press

Wyma KD (1997) Moral responsibility and leeway for action. American Philosophical Quarterly 34(1):57–70