[DRAFT] Frankfurt paper

Sander Beckers

1 Introduction

Note: the idea is to use a formal framework for expressing causal relations, and start from a very general definition of responsibility. Then I look at the Frankfurt cases, and use them as a guide to filling in the details of the definition. The first part that will be filled in is the causal relation that is required for responsibility. Then I state a general version of the Principle of Alternative Possibilities, and see how that can be satisfied. This will result in the relevant alternative possibility for the agent not to be responsible being grounded in the fact that the agent believes there to be an alternative possibility for the causal relation between his choices and the outcome.

2 Structural Equations Modelling

We briefly introduce a simple version of structural equations modelling, which is the most popular formal language used to represent causal models. In general, structural equations allow functional dependencies between continuous variables, or discrete variables with possibly an infinite domain. However we restrict attention to examples made up of discrete variables with a finite domain, and propositional formulas. Further, in the majority of cases the variables are Boolean. This is why we restrict attention to those kinds of models. For a detailed introduction, see (Pearl, 2000).

Note: I will generalise to multi-valued variables and functional dependencies of any kind.

A structural model consists of a set of *endogenous* variables $\vec{V}$, a set of *exogenous* variables $\vec{U}$, and a causal model M. Although we only consider models with Boolean variables, we should point out that the results we will present can easily be generalized to allow for multi-valued variables as well. We explain this below.

A model M is a set of *structural equations* so that there is exactly one equation for each variable $V_i \in \vec{V}$. An equation takes the form $V_i := \phi$, where ϕ is a propositional formula over $\vec{V} \cup \vec{U}$. For any variable V_i, we denote by ϕ_{V_i} the formula in the equation for V_i in M. We follow the customary practice of leaving the equations for variables that depend directly on the exogenous variables implicit, and simply state the value they take in each particular story.

For an assignment $(\vec{v}, \vec{u})$ of values to the variables in $\vec{V} \cup \vec{U}$, we denote by $\phi^{(\vec{v},\vec{u})}$ the truth value obtained by filling in the truth values $(\vec{v}, \vec{u})$ in the formula ϕ. An assignment $(\vec{v}, \vec{u})$ *respects* M, if for each endogenous variable V_i, its value $v_i = \phi_{V_i}^{(\vec{v},\vec{u})}$. As usual, we only consider models M in which the equations are acyclic, which implies that for each assignment $\vec{u}$ to $\vec{U}$, there is exactly one assignment $(\vec{v}, \vec{u})$ that respects M. Therefore, we refer to $\vec{U} = \vec{u}$ as a *context*. For every value $\vec{u}$ of $\vec{U}$, we call the pair $(M, \vec{u})$ a *causal setting*. We write $(M, \vec{u}) \models \phi$ if $\phi^{(\vec{v},\vec{u})} = \mathbf{true}$ for the unique assignment $(\vec{v}, \vec{u})$ that respects M.

A *literal* L is a formula of the form $V_i = v_i$ or $U_i = u_i$. Our restriction to Boolean variables is made concrete here: the only values v_i we consider are $\mathbf{true}$ and $\mathbf{false}$. Hence our definitions and results can be generalised by simply lifting this restriction. (See the Appendix for some more details.)

We will use the atom V_i as a shorthand for $V_i = \mathbf{true}$, and the negated atom $\neg V_i$ as a shorthand for $V_i = \mathbf{false}$. If V_i is endogenous, we write ϕ_{L_i} for ϕ_{V_i} in both cases.

A causal model M is a tool to represent *counterfactual* relations between variables, in the sense that changing the values of the variables on the right-side of an equation can change the value of the variable on the left-side, but not vice versa. This makes them suitable devices to model *interventions* on an actual setting, meaning changes to the value of a variable V_i that affect only the values of variables that depend on V_i, but not those on whom V_i itself depends.

Syntactically, we make use of the $do()$-operator introduced by Pearl (2000) to represent such an intervention. For a model M and an endogenous variable V_i, we denote by $M_{do(V_i)}$ and $M_{do(\neg V_i)}$ the models that are identical to M except that the equations for V_i are $V_i := \mathbf{true}$ and $V_i :=$

false, respectively. Hence for a causal setting $(M, \vec{u})$ such that $(M, \vec{u}) \models C$, the causal setting $(M_{do(\neg C)}, \vec{u})$ corresponds to the counterfactual setting resulting from the intervention on $(M, \vec{u})$ that prevents C.

2.1 Responsibility Settings

We define some set of exogenous variables $\vec{C} \subseteq \vec{U}$ to be the set of *agent variables*. These are variables which are under "direct control" of the agent, where we leave open what this means precisely. The reason is that we do not want to make any assumptions about the nature of free will, or agent causation. Instead we simply accept that in some way or other the agent always has control over the values that these variables take. For every $C_i \in \vec{C}$, we say that the agent *directly chooses* C_i w.r.t. M. $\vec{W}$ denotes the exogenous variables that are not under the agent's control, i.e., $\vec{W} = \vec{U} \setminus \vec{C}$.

In addition to the objective structural model, there is also the agent's structural model, i.e., the structural model that the agent believes to be correct. We use M, $\vec{V}$, and $\vec{U}$ to refer to the objective structural model, whereas the agent's structural model consists of M_a, $\vec{V_a}$, and $\vec{U_a}$.

Further, we assume that the agent has a probability distribution $P(\vec{W_a})$ over the exogenous variables that are not under his control. (This could be generalised to a probability distribution over causal models as well.)

We assume that responsibility judgments are made relative to a *responsibility setting* $\mathcal{R}$, which is the combination of an objective causal setting $(M, \vec{u})$ and the agent's (probabilistic) causal setting $(M_a, P(\vec{W_a}))$. In other words, $\mathcal{R} = \{(M, \vec{u}); (M_a, P(\vec{W_a}))\}$.

3 A General Definition of Responsibility

We present a general definition of responsibility that captures a commonly accepted structure of what a definition of responsibility should look like. We assume that responsibility for an outcome O is determined by there being some agent variable A that stands in a particular relation to the outcome, namely a relation that grounds the responsibility for O in the agent's responsibility for A. The purpose of this paper is to fill in the details of this relation along the way, by looking at the restrictions that the Frankfurt cases impose on the various conditions.

An important restriction throughout this paper is that we assume there to be only one morally relevant outcome O. In this manner we can ignore issues relating to there being some other important outcome O' which plays a role in understanding the agent's relation to O. Therefore in our setting the notion of responsibility is equivalent to the notions of blame- and praiseworthiness: if O is a "morally negative" outcome, then being responsible for O is equivalent to being blameworthy for O, and likewise for a positive outcome and being praiseworthy.[1]

Also, we ignore responsibility due to an epistemic failing, i.e., we ignore cases where an agent is responsible for some outcome due to him having incorrect beliefs about the causal model that he should not have had. (Eg., a doctor whose operation results in the patient's death because he had failed to properly inform himself about all the relevant details.)

Note: I will motivate this definition by referring to Braham and van Hees (2012) and work by Sartorio. Suggestions on other sources to cite are welcome.

Definition 1 (Responsibility)**.** *Given a responsibility setting $\mathcal{R}$ such that $(M, \vec{u}) \models O$, the agent is responsible for the outcome O if the agent directly chose O or there exists a literal A satisfying the following conditions:*

- **(Agency Condition)** *The agent directly chose A.*
- **(Causal Condition)** A causally contributes *to O w.r.t. $(M, \vec{u})$.*
- **(Doxastic Condition)** *The agent believes X.*

Agency Condition (AC): Braham and van Hees assume that all actions performed by the agent are intentional, without specifying what this means. Our condition is similar, except that we generalise this assumption to whatever variables are directly chosen by the agent, be they actions or not.

Causal Condition (CC): For now we leave it entirely open how to define what it means to causally contribute. Our aim is precisely to use the Frankfurt cases to guide us in figuring out what a good definition of causal contribution should be. (Obviously in some way or other this will have to be related to actual causation.)

Doxastic Condition (DC): A responsibility judgment is based both on facts about the objective causal relations between events/omissions, and on facts about the agent. Concretely, for an agent to be responsible it is necessary that the agent holds an appropriate set of beliefs upon which she acts, or makes choices. As with the Causal Condition, we start out without making any assumptions about this condition.

The first two conditions are entirely similar to those used by Braham and van Hees. They add as a third condition that "The agent should have had a reasonable opportunity to have done otherwise", which is a version of the Principle of Alternative Possibilities. Instead of starting out with such a condition, our goal is to see what room – if any – the Frankfurt cases leave for such a condition to be accepted.

[1] In terms of Braham and van Hees (2012), this restriction is equivalent to assuming that all strategies are *eligible*.

4 Frankfurt Cases

Note: For now I limit myself to giving two concrete Frankfurt cases, and restrict my analysis to those. However I think it might be useful to also formally define what a Frankfurt case is, so that my results are more general. On the other hand, a formal version will be tedious and might attract a smaller audience, so maybe that's something for an Appendix only? The cases I will consider are the two extremes: the original Frankfurt case, and the most sophisticated one out there, which is the latest one by Pereboom (I think?).

A crucial part of any Frankfurt case is that the agent's beliefs are identical to those in the non-Frankfurt version of the case. Therefore for now we can simply ignore the Doxastic Condition entirely.

4.1 The Original Frankfurt Case

Here's the original Frankfurt case, from Frankfurt (1969):

Example 1 (Original)**.** *Suppose someone – Black, let us say – wants Jones to perform a certain action. Black is prepared to go to considerable lengths to get his way, but he prefers to avoid showing his hand unnecessarily. So he waits until Jones is about to make up his mind what to do, and he does nothing unless it is clear to him (Black is an excellent judge of such things) that Jones is going to decide to do something* other *than what he wants him to do. If it does become clear that Jones is going to decide to do something else, Black takes effective steps to ensure that Jones decides to do, and that he does do, what he wants him to do. Whatever Jones' initial preferences and inclinations, then, Black will have his way.*

The following causal model captures the background of this story, where $Motivated$ is an agent variable.

$$\begin{aligned} Action &:= Decide. \\ Decide &:= Motivated \vee Black. \\ Black &:= Sign. \\ Sign &:= \neg Motivated. \end{aligned}$$

Note: Explain variables. The context is such that $Motivated$ holds.

We take the non-Frankfurt version of *Original* to be the version of the example in which the strange Frankfurt intervener is left out, i.e., the following model, with the same context:

$$\begin{aligned} Action &:= Decide. \\ Decide &:= Motivated. \\ Sign &:= \neg Motivated. \end{aligned}$$

We refer to the non-Frankfurt version of *Original* as $Original^*$.

Intuitively everyone would agree that Jones is responsible for $Action$ in $Original^*$, and for most people this intuition carries over to *Original*. We take Frankfurt's argument to be that a good definition of responsibility should respect these intuitions:

Principle 1 (Frankfurt Principle)**.** *If you a have definition of responsibility such that Jones is responsible for $Action$ in $Original^*$, then your definition should also judge Jones to be responsible for $Action$ in $Original$.*

We aim to analyse the implications of accepting the Frankfurt Principle. So we start with the intuitive assumption that Jones is responsible for $Action$. Since $Motivated$ is the only agent variable, this implies that $Motivated$ produced $Action$ in $Original^*$. Since the entire causal influence of $Motivated$ on $Action$ is mediated through $Decide$, it is safe to assume that $Motivated$ producing $Decide$ is entirely dependent on whether or not $Motivated$ produces $Decide$.

To put this into in perspective, it is helpful to introduce the following familiar example from the literature on actual causation, which is a case of *Early Preemption* from Hitchcock (2001):

Example 2 (Backup)**.** *An assassin-in-traininig is on his first mission. Trainee is an excellent shot: if he shoots his gun, the bullet will fell Victim. Supervisor is also present, in case Trainee has a last minute loss of nerve (a common affliction among student assassins) and fails to pull the trigger. If Trainee does not shoot, Supervisor will shoot Victim herself. In fact, Trainee performs admirably, firing his gun and killing Victim.*

The following is an appropriate model for this story, where the context is such that $Trainee$ holds:

$$\begin{aligned} Victim &:= Trainee \vee Supervisor. \\ Supervisor &:= \neg Trainee. \end{aligned}$$

The causal relation between $Motivated$ and $Decide$ in *Original* is almost structurally identical to the causal relation between $Trainee$ and $Victim$ in *Backup*. Similarly, the causal relation between $Motivated$ and $Decide$ in $Original^*$ is structurally identical to that between $Trainee$ and $Victim$ in the extremely simple one equation example $Victim := Trainee$.

Hence all that the Frankfurt Principle states, is that if $Trainee$ is a causal contributor to $Victim$ in this simple example, then $Trainee$ should also be a causal contributor to $Victim$ in *Backup*. Given that most authors consider $Trainee$ to be an actual cause of $Victim$ in both *Backup* and in the simple example, this is an easy challenge to meet: simply define causal contribution as actual causation is defined by any of the following authors: (Hitchcock, 2001; Woodward, 2003; Hall, 2004, 2007; Halpern

and Pearl, 2005; Halpern, 2016; Weslake, 2015). Another option is to define causal contribution as *production*, a concept that we have formally defined in (Beckers and Vennekens, 2016) as a fundamental building block to defining actual causation.

Braham and van Hees (2012) use the NESS test to fill in the Causal Condition, in which case the Frankfurt Principle is not satisfied: in *Original* we have that $\emptyset$ is a sufficient set for $Action$, which implies that $Motivated$ does not meet the NESS test for $Action$. They do apply their approach successfully to an example that looks like a Frankfurt case, but in fact it is not. Their example allows a situation in which Black changes his mind and does not intervene even if Jones is not motivated to perform Action. Since it is an essential feature of any Frankfurt case that one assumes the backup mechanism to be infallible, their example is ill-chosen.

Note: The causal condition used by Sartorio also fails: reference to the "All Roads lead to Rome" example as an illustration.

4.2 Pereboom's Frankfurt Case

One might object that our analysis is too simple, because by now there are far more complex Frankfurt cases than *Original*. 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.

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

Note: Explain variables. $Attention$ and C are agent variables. The context is such that $\neg Attention$ holds, and C is undetermined.

$$\begin{aligned} Tax &:= Decide. \\ Decide &:= \neg Attention \vee Device. \\ Device &:= Attention. \end{aligned}$$

The model for the version without the Frankfurt intervener, i.e., a model for *Tax* $Evasion^*$:

$$\begin{aligned} Tax &:= Decide. \\ Decide &:= \neg Attention \vee C. \end{aligned}$$

In this example the causal relation of importance is that between $\neg Attention$ and $Decide$. As with the *Original* example, defining causal contribution as either actual causation along the lines of the definitions mentioned above, or defining it as production, gives the desired result that $\neg Attention$ causally contributes to $Decide$ in both *Tax Evasion* and *Tax* $Evasion^*$.

4.3 Reverse Frankfurt Case

Note: This is an example that shows we should not use the HP-definitions, or any of the other definitions of causation, and hence we have to take my notion of production.

Example 4 (Injection)**.** *Jones is standing next to the hospital bed of Patient, with a syringe in his hands. Patient suffers from a rare lethal disease, and is about to die. The syringe contains a medicine for Patient's condition, but unfortunately Patient is allergic to the medicine. In fact, if Jones were to inject the medicine, Patient would die from an allergic Reaction. Jones knows all of this, except for the fact that Patient is suffering from the lethal disease. In other words, Jones believes that Patient will die only if he injects the medicine. Since Jones dislikes Patient very much, he injects the medicine and Patient dies from the allergic Reaction.*

Model (with $Inject$ an agent variable):

$$\begin{aligned} Patient &:= Reaction \vee \neg Inject. \\ Reaction &:= Inject. \end{aligned}$$

As with the Frankfurt cases, intuitively Jones is responsible for Patient's death. What do we get applying our definition?

Jones is responsible for $Inject$. Assuming that DC is also fulfilled, his responsibility depends on CC: does $Inject$ causally contribute to $Patient$?

According to just about every definition of actual causation that there is, $Inject$ did not cause $Patient$. This suggests that it's a bad idea to define causal contribution as actual causation.

However things get even worse. Imagine that unbeknownst to Jones, Black is standing behind the curtains, watching Jones' every move. If it were to become clear that Jones would not inject the medicine, Black would shoot Patient. In other words, we change our example into a Frankfurt case. Now the model becomes:

$$Patient := Reaction \vee Black.$$
$$Reaction := Inject.$$
$$Black := \neg Inject.$$

This model exhibits what is known as switching behaviour: $Inject$ acts like a switch, as it determines which of two mechanisms is initiated that both have the same effect. In this model the HP-definitions (and many others as well) *do* judge $Inject$ to be a cause of $Patient$. So if we were to define causal contribution as actual causation along the lines of the HP-definitions, we end up with the paradoxical result that adding a Frankfurt intervener can turn a non-responsible agent into a responsible agent!

Fortunately our definition of production does not suffer from this defect, as it judges $Inject$ to be a producer of $Patient$ in both cases, in agreement with our intuitions that Jones is responsible for Patient's death in both cases.

Note: something about: maybe someone would argue that the first model is inappropriate, but that sounds rather ad hoc. The key insight is that the possible counterfactual stories shouldn't really matter for responsibility at all. Since counterfactuals do matter for actual causation, the causal condition for responsibility is not identical to causation.

Conclusion of this section: the lesson to be learned from the Frankfurt cases is that the causal condition for responsibility in Definition 1 should be filled in with our notion of production.

Note: here I will say something about the Actual Sequence view of free will, and how production is a very natural way to cash out the causal part of that view: all that causally matters for responsibility is the actual sequence that produced the outcome, regardless of any counterfactuals.

5 Principle of Alternative Possibilities

The PAP as stated by Frankfurt can be expressed in our framework as follows:

Principle 2 (Naive PAP)**.** *Given a responsibility setting $\mathcal{R}$ such that $(M, \vec{w} \cup \vec{c}) \models O$, if the agent is responsible for the outcome O and O is an action by the agent, then there exist values $\vec{c'}$ of the agent variables such that $(M, \vec{w} \cup \vec{c'}) \models \neg O$.*

The Frankfurt cases convincingly show that Naive PAP is false. However those cases also reveal two key insights as to why this principle fails.

First, it singles out the action of an agent as fulfilling a special role compared to other elements in the causal chain. This is based on the mistaken assumption that performing an action intentionally is equivalent to performing an action that one could have chosen not to perform.

Second, it assumes that an alternative possibility with regards to O has to be an alternative possibility in which O does not occur. Concretely, there being an alternative possibility is assumed to be identical to O being counterfactually dependent on choices made by the agent. As has been amply shown in the causation literature, counterfactual dependence is only one form of difference making, one which is so strong that it applies only to a very limited set of examples.

Instead, a more subtle way of making a difference presents itself by focussing not on whether the outcome O depends on an event A, but on whether the relation of interest that holds between A and O also holds for all alternatives of A. Since in our case the relation of interest between A and O is that of grounding responsibility, we get the following more subtle version of PAP:

Principle 3 (PAP)**.** *Given a responsibility setting $\mathcal{R}$ such that $(M, \vec{w} \cup \vec{c}) \models O$, if the agent is responsible for the outcome O then there exist values $\vec{c'}$ of the agent variables such that the agent would not be responsible for O in $(M, \vec{w} \cup \vec{c'})$.*

We now consider how we can further fill in the details of our definition of responsibility (Definition 1) in a way that ensures PAP to be satisfied.

Looking at $Original^*$, we see that $Action$ is counterfactually dependent on $Motivated$. Hence PAP is trivially satisfied for this example, since changing $Motivated$ into $\neg Motivated$ results in $\neg Action$. This is no longer true in *Original*, since $Black$ ensures that Jones performs $Action$ even if he is not motivated to do so. In fact, changing $Motivated$ into $\neg Motivated$ in *Original* has no effect on the Causal Condition: in both cases the value of $Motivated$ will be a producer of $Action$.

If we step back for a second and consider actual causation as defined by Halpern and Pearl (2005) or Halpern (2016), we can make the following observation: $Motivated$ does cause $Action$ in *Original*, but $\neg Motivated$ would not be an actual cause of $Action$ in *Original*. Therefore it might be tempting to retrace our steps, and define causal contribution as actual causation after all. However, as we have seen with the Reverse Frankfurt case, these definitions make it very easy to turn a non-cause into a cause in these types of examples: all one has to do is add an intermediate variable V in between $Motivated$ and $Decide$, i.e., change the

equation for $Decide$ into $Decide := V \vee Black$, and add $V := Motivated$, and $\neg Motivated$ turns into a cause of $Action$. Because of this extreme sensitivity to modelling details, we do not consider this a viable strategy to pursue.

Instead we draw the conclusion that the difference between the actual responsibility setting in which the agent chooses $Motivated$, and the counterfactual responsibility setting in which the agent chooses $\neg Motivated$, is not to be found in any difference between the *objective* causal relations. In other words, taking the Frankfurt cases seriously means accepting the fact that responsibility for O does not require there to be an alternative possibility such that the *objective* causal relations between O and the agent's choices are any different than in the actual story.

The obvious conclusion is that we should shift focus to the *doxastic* causal relations, i.e., the causal relations as seen from the agent's perspective. Concretely, rather than assuming that the validity of PAP is due to there always existing an alternative possibility for which CC is not satisfied, we should base it on the claim that there always exists an alternative possibility for which *the agent believes that CC is not satisfied*.

Applying this idea to the Frankfurt cases confirms that this solution offers a way to preserve PAP. Since Jones is oblivious to the existence of $Black$, the agent's causal model for *Original* is the objective causal model for $Original^*$. Therefore Jones believes that $\neg Motivated$ would not have produced $Action$. In other words, he believes that there is an alternative possibility for which CC is not satisfied, implying that he believes there to be an alternative possibility in which he would not be responsible for $Action$.

Likewise, the agent's causal model in *Tax Evasion* is the objective causal model for $TaxEvasion^*$, according to which choosing $Attention$ and $\neg C$ results in there being no producer of Tax for which the agent is responsible.

This leads to a first attempt at formulating a suitable Doxastic Condition: the agent believes both that one of her actual choices A will produce O, and that she can choose the other agent variables such that $\neg A$ would not produce O.

Note however that we have formulated the Doxastic Condition in terms of an agent's full beliefs, i.e., beliefs with probability 1. Recalling our assumption from Section 2 that the agent has a probability distribution $P(\vec{W}_a)$ over the exogenous variables that are not under her control, we should refine our condition. In the examples considered so far the agent's uncertainty could be ignored, since it held that $\vec{W}_a = \emptyset$. However, in general it is clearly too strong to demand that the agent believes her choices will certainly produce the outcome, since this would deny her responsibility whenever she is uncertain about the consequences of her choices.

Imagine for example that Jones knows he is not much of a sharp shooter, and he believes that if he were to shoot at Victim, he may very well miss. As it turns out, Jones' shot is accurate, and Victim dies. The mere fact that he believed he might have missed does not imply that he is not responsible for killing his victim.

Therefore our Doxastic Condition should focus on what the agent *expects* her choices to produce, as follows: the agent believes that the actual values $\vec{c}$ of the agent variables are more likely to produce O than any alternative setting $\vec{c'}$ of the agent variables. We can interpret this condition as stating that a necessary condition for responsibility is that the agent *tries to become responsible*.

Note: refer back to Braham and van Hees's condition, to show that it is in the same spirit. Also refer to Sartorio's transition principle, which is also similarly motivated.

6 Putting it all Together

Definition 2 (Responsibility)**.** *Given a responsibility setting $\mathcal{R}$ such that $(M, \vec{w} \cup \vec{c}) \models O$, the agent is responsible for the outcome O if the agent directly chose O or there exists a literal A satisfying the following conditions:*

- **(Agency Condition)** *The agent directly chose A.*
- **(Causal Condition)** *A is a* producer *of O w.r.t. $(M, \vec{w} \cup \vec{c})$.*
- **(Doxastic Condition)** $\vec{c} = \arg\max_{\vec{c}} P(\{\vec{w_a} | \exists c_i \in \vec{c} :$ *c_i is a producer of O w.r.t. $(M_a, \vec{w_a} \cup \vec{c})\}$).*

Note: maybe this version of the Doxastic Condition is overly complicated: for most examples it suffices to say: the agent believes that A is more likely to produce O than $\neg A$. Maybe I should use that version, and leave the complicated one for the Appendix? Also: strictly speaking the current version leaves open the possibility that all choices $\vec{c}$ result in equal probabilities, in which case PAP is false. So we should add that this may not be, but that makes it even more tedious.

As we saw in *Tax Evasion*, it is possible that some agent variable C_i is undetermined in the actual story. Given that the agent believes she could have chosen any value she liked, we read Definition 2 as stating that there should exist at least one value c_i for which DC holds.

6.1 Illustrations

Note: here I will give some more examples to illustrate the intuitiveness of my definition.

6.2 Robust Alternatives

Opponents of the PAP do not deny that there always exist alternative possibilities which the agent could have chosen.

However they claim that those alternatives are not always *robust* enough to ensure that the agent certainly has an alternative possibility available to her which, had she chosen it, would have precluded her from being responsible for the outcome. Given that our definition assumes there to be at least one alternative setting $\vec{c'}$ that does not maximise the probability of producing the outcome, our definition does ensure the existence of a robust alternative, and hence their claim is false.

Since Pereboom (2009) motivates his definition of robustness using *Tax Evasion*, it will be helpful to compare his analysis to ours. He defines an alternative to be robust when:

> she [the agent] 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.

The reason Pereboom believes that this condition is not met for Joe in *Tax Evasion*, is that "Joe does not understand, and, moreover, he has no reason to believe, that voluntarily achieving the requisite level of attentiveness would or would likely preclude him from responsibility for choosing to evade taxes". Applying our definition of responsibility this is translated into stating that Joe does not believe $Attention$ would be less likely to produce Tax than $\neg Attention$ is. We offer two replies.

First of all, it is crucial for Pereboom to focus solely on Joe's choice regarding $Attention$ and to ignore his choice regarding C. However he considers ignoring C to be entirely justified, since there exists no alternative possibility in which Joe will ever get to choose a value for C, despite Joe believing that there is. We do not find this convincing.

Imagine that C were not an agent variable, but an exogenous variable for which $P(C) \neq 1$. In that case, Joe definitely believes that choosing $Attention$ would preclude him from responsibility for Tax, since he believes it would maximise the probability that Tax would not occur. But of course Joe is wrong in believing this, since there exists no alternative possibility in which C will ever get a value. If him being wrong about this does not justify ignoring the possible values that Joe believes C could take on, then why should it do so when C is an agent variable? What reason is there to treat these situations differently? But if we do consider Joe's beliefs regarding C to be relevant, then we observe that Joe believes choosing $Attention$ and $\neg C$ would definitely preclude him from being responsible for Tax, as he believes Tax wouldn't even occur.

Second of all, and more importantly, Joe in fact *does* believe that $Attention$ would be less likely to produce Tax than $\neg Attention$ is, and hence on our view of responsibility he *does* have a reason to believe that "that voluntarily achieving the requisite level of attentiveness would or would likely preclude him from responsibility for choosing to evade taxes".

Joe's causal model is the objective causal model for *Tax* $Evasion^*$. According to that model, $\neg Attention$ certainly produces Tax, whereas $Attention$ would not produce Tax in case $\neg C$ holds. Since the story implies that Joe does not start out believing that he will certainly evade taxes, he believes there is a positive probability that he would choose $\neg C$ if it were to come to it. Now of course if Joe had chosen $Attention$, the Frankfurt intervener would have prevented Joe to make any choice regarding C. However Joe does not know that, and hence he believes $Attention$ to be a less likely producer of Tax than $\neg Attention$.

Side Conclusion: there are two camps: those who claim that PAP is a necesssary condition, and those who claim that facts about the Actual Sequence are sufficient. My view satisfies both claims, and hence offers a natural compromise away from the stalemate.

7 Conclusion

References

Beckers S, Vennekens J (2016) A principled approach to defining actual causation. Synthese forthcoming

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

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

Hall N (2004) Two concepts of causation. In: Collins J, Hall N, Paul LA (eds) Causation and Counterfactuals, The MIT Press, pp 225–276

Hall N (2007) Structural equations and causation. Philosophical Studies 132(1):109–136

Halpern J (2016) Actual Causality. MIT Press

Halpern J, Pearl J (2005) Causes and explanations: A structural-model approach. part I: Causes. The British Journal for the Philosophy of Science 56(4):843–87

Hitchcock C (2001) The intransitivity of causation revealed in equations and graphs. Journal of Philosophy 98:273–299

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

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

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