Causal Models and the Principle of Alternative Possibilities

Sander Beckers

Department of Philosophy and Religious Studies
Utrecht University
sanderbeckers.com

Causes, Norms, Decisions, Hannover, August 16, 2018

Utrecht University

Outline

Outline

Context of the Frankfurt debate

- What is Free Will? When are we morally responsible?
- Metaphysical debate: Determinism vs indeterminism, compatibilism vs incompatibilism, source vs leeway, ...
- Practical debate: Frankfurt cases vs Principle of Alternative Possibilities (PAP)
- Formal debate: $\emptyset$
  - Exceptions: (Braham & Van Hees, ...?)

Analysis of the Frankfurt debate

- Problem: Almost half a century with little progress and plenty of confusion. Eg. "stalemate" (Fischer, 1994, Timpe 2006)

- Solution: Causality is key!
  - "This does, to be sure, suppose that there is some sort of causal relation between Jones's state at the time of the twitch and his subsequent states." (Frankfurt, 1969).

  - Fischer 1999, overview paper, 48 pages:
    - "causal": 83 occurrences
    - "causation": 23 occurrences
    - "cause": 24 occurrences

Analysis of the debate (cont.)

- Pereboom 2009:
  - "I ... endorse instead a type [of incompatibilism] that ascribes the most significant explanatory role to the nature of the causal history of the agent's production of the action."
  - "Its distinguishing features are these: ... the absence of the cue for intervention in no sense causally determines the action the agent actually performs."
  - "However, Joe's imagining in this way being punished is not causally sufficient for his failing to choose to evade taxes.

My claims:

1. Causal modelling of Frankfurt cases is essential to move forward.

2. All Frankfurt cases are entirely compatible with certain - interesting - versions of the PAP (but not with others).

Outline

1969: Harry Frankfurt "Alternate Possibilities and Moral Responsibility"

- Principle (PAP): responsibility $\Rightarrow$ alternative possibilities
- Intuition: agent is responsible in *Frankfurt case*
- Claim: there are no alternative possibilities in *Frankfurt case*
- Conclusion: principle is false
- General conclusion: need more nuanced relationship between Responsibility and PAP

1973: David Lewis "Causation"

- Principle: causation $\Rightarrow$ counterfactual dependence
- Intuition: there is causation in *Early Preemption* example
- Claim: there is no counterfactual dependence in *Early Preemption*
- Conclusion: principle is false
- General conclusion: need more nuanced relationship between Causation and Counterfactual Dependence

Counterfactual approach to causation

- Formal tools: Structural Equations Modelling (Pearl (2000); Causal Bayesian Networks (Spirtes et al., 2000), CP-logic (Vennekens et al., 2009)
- Progress in definitions of causation: (Hitchcock, Hall, Woodward, Yablo, Halpern & Pearl, Weslake, Beckers & Vennekens, Halpern,...)
- Dozens of interesting examples and discussions on how to model them.

Outline

Strong PAP

Principle 1 (**Strong PAP**)

If an agent is non-derivatively responsible for performing an act A, then she could have voluntarily not performed A.

Definition 1

Given a causal setting $(M, \vec{u})$ such that $(M, \vec{u}) \models C \wedge E$, E is *counterfactually dependent* on C if $(M_{do(\neg C)}, \vec{u}) \models \neg E$.

Principle 2 (**Dependence**)

If E is dependent on C in a causal setting $(M, \vec{u})$, then C is a cause of E w.r.t. $(M, \vec{u})$.

Derivative Responsibility

(Widerker, 2006):

"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."

Weak PAP

Principle 3 (**Weak PAP**)

If an agent is responsible for performing an act A, then she could have voluntarily chosen an alternative such that she would not have been responsible for A.

Principle 4 (**Asymmetry**)

If C is a cause of E w.r.t. $(M, \vec{u})$, then $\neg C$ is not a cause of E w.r.t. $(M_{do(\neg C)}, \vec{u})$.

Weak vs Strong PAP

Is **Weak PAP** really weaker than **Strong PAP**?

- Scope now also includes derivative responsibility. But: assume responsibility for A is derived from responsibility for B. Then avoiding responsibility for B presumably implies avoiding responsibility for A too.

- Ignoring scope: $\neg A$ implies not responsible for A, but not vice versa.

So **Strong PAP** (typically) implies **Weak PAP**.

PAP in the literature

Defenders of only **Weak PAP** (or variants thereof): (McKenna, 1997; Wyma 1997; Otsuka, 1998; Braham and Van Hees, 2012).

Those who use the intuitive appeal of **Weak PAP** to justify **Strong PAP**: (Ginet, 1996; Widerker, 2000; Palmer, 2014).

But **Strong PAP** is much more vulnerable to Frankfurt cases than **Weak PAP**!

Moderate PAP

Principle 5 (**Moderate PAP**)

If an agent is responsible for performing an act A, then she could have chosen a robust alternative such that she would not have been responsible for A.

Robust = voluntary + ...?

Outline

Example: Voting (Fischer, 1982, 1999, 2010)

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).

Example (cont.): Voting

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.

Argument against PAP

Assume first that there is no device. Then this example is 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 device, 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.

Hence the ability to do otherwise is not a necessary condition for responsibility.

Counterargument

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. So Jones had the alternative possibility of exhibiting a different sign. (So-called "flicker of freedom".)

Rebuttal to the Counterargument

A flicker of freedom, as the name suggests, is not a genuine alternative possibility. Fischer (1999): "The power involuntarily to exhibit a different sign seems to me to be insufficiently robust to ground our attributions of moral responsibility"

My claim: Causal model shows that this is besides the point.

Observation

- The device *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 ($Choice = 1$), then Jones does in fact vote for Gore ($Vote = 1$), and vice versa.
- That is, $Vote = 1$ is counterfactually dependent on $Choice = 1$.
- This means that Jones's act ($Vote = 1$) is *completely determined* by his choice ($Choice = 1$).

Observation (cont.)

Recall derivative responsibility: "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 responsibility for $Vote = 1$ is derived from responsibility for $Choice = 1$, then **Strong PAP** doesn't even apply.

Plausible reply: focus on responsibility for $Choice = 1$ and the argument still works.

Building the causal model: variables

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 ($Sign = 1$), and Black's device not being triggered ($Device = 0$).

To each 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 ($Device = 1$).

Typical discussions of this example (and many others like it) assume these variables are all we need.

Try to build the equations, and you see this is a mistake.

Building the causal model: equations

SEMs (Pearl, 2000): the mechanism that causally determines a variable X is captured by an equation of the form $X := f(\vec{Y})$.

Vote is counterfactually dependent on *Choice*, so we get:
$\mathit{Vote} := \mathit{Choice}$

The equation for *Choice* forms the most crucial part of the causal model.

Start with normal setting (=no device):

- Jones's choice (*Choice*) is perfectly correlated with him exhibiting a particular involuntary sign: if $\mathit{Sign} = 1$, we have $\mathit{Choice} = 1$, and if $\mathit{Sign} = 0$, we have $\mathit{Choice} = 0$.

Perfect correlation $\Rightarrow$ one causes the other, or common cause.

Building the causal model: equations

Choice causes *Sign*? Obviously not. (*Choice* is at T_2, *Sign* is at T_1.)

Sign causes *Choice*? Not the intended meaning of "sign". Also, it's exhibited involuntary, implying that it is merely a side-effect of Jones's deliberation.

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.

Let us call this event $Reflection = 1$ (" After serious reflection, he chooses to vote for Gore and does vote for Gore ".)

$Sign := Reflection$ and $Choice := Reflection$.

Causal model: *Dependence*

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

By assumption, **Strong PAP** is confirmed here. Hence, Jones could have voluntarily voted differently ($\textit{Vote} = 0$). Given the model, this means Jones could have voluntarily reflected differently ($\textit{Reflection} = 0$)

What is "voluntary control"?

- 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 voluntary control over Reflection*.

Causal model with device: *Early Preemption*

$Device := \neg Sign$.

Equations for *Vote* and *Sign* are unaffected.

Only *Choice* is influenced by the device: 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.

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

Causal model for Voting: *Early Preemption*

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

The relation between $\mathit{Reflection} = 1$ and $\mathit{Choice} = 1$ is just *Early Preemption*.

So as far as the causal properties are concerned, going from a normal case to a Frankfurt case means going from *Dependence* to *Early Preemption*.

Analysis

Normal case: Jones had control over *Choice* because he had control over *Reflection*.

How does the addition of the device change this? Jones loses control over *Choice*, while retaining his control over *Reflection*.

- **Strong PAP**: If responsibility for $Choice = 1$ is non-derivative, then **Strong PAP** is falsified. If responsibility for $Choice = 1$ is derived from Jones's responsibility for $Reflection = 1$, then it falls beyond the scope of **Strong PAP**.
- **Weak PAP**: Jones had control over Reflection. If *Reflection* had been 0, then he would have been *forced* by the device to choose to vote for Gore, rather than make that choice voluntarily. Wide consensus: if forced, then not responsible. So there existed an alternative possibility such that he voluntarily could have avoided being responsible for $Choice = 1$, confirming **Weak PAP**.

Example: Voting

Conclusion:

1. 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.
2. If $Choice = 1$ is a case of non-derivative responsibility, then **Strong PAP** is falsified, but **Weak PAP** is not.
3. If $Choice = 1$ is a case of derivative responsibility, then neither **Strong PAP** nor **Weak PAP** are falsified by the Frankfurt argument.

Possible objection: your model conflicts with my metaphysical views on free will and determinism.

Reply: fair enough, let's (finally) have a discussion about causal models then.

Outline

Example: Tax Evasion

Without device: *Probabilistic Early Preemption*

$$\mathit{Choice} := \neg \mathit{Attention} \vee \mathit{Override}.$$
$$\mathit{Override} := \mathit{Attention} \wedge \mathit{Self}\,.$$

With device: *Early Preemption*

$$\mathit{Choice} := \neg \mathit{Attention} \vee \mathit{Device}.$$
$$\mathit{Device} := \mathit{Attention}.$$

So as far as the causal properties are concerned, going from a normal case to a Pereboom-Frankfurt case means going from *Probabilistic Early Preemption* to *Early Preemption*.

Example: Tax Evasion

Given that **Strong PAP** and **Weak PAP** only speak about about the properties of the post-device causal model, analysis is identical to that for Voting.

What changes? The agent's *beliefs* about the causal model.

Moderate PAP

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 **Moderate PAP**:

Principle 6 (**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.*

Pereboom's robustness (2009)

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.

Pereboom's robustness (2009)

Motivation: 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 that mere existence of an alternative does not suffice: the agent "would *likely to be* precluded from the responsibility she actually has".

Pereboom's robustness (2009)

Why "likely"? Pereboom:

- Existence of alternative is too weak: there always exists "some" alternative, as long as the agent is a healthy sceptic.
- Certainty is too strong: "likely" sounds good enough.

"the threshold probability, as one would expect, is difficult or impossible to determine", and therefore he settles for "likely".

An (obvious?) improvement is possible.

Obvious threshold probability

The agent is comparing two alternatives: $Attention = 0$ vs $Attention = 1$.

Assume that in the actual scenario, $Choice = 1$. When is he "off the hook", i.e., not responsible for $Choice = 1$?

Just compare $P(\text{Responsible for } Choice = 1 | Attention = 0)$ to $P(\text{Responsible for } Choice = 1 | Attention = 1)$: if the agent "willed" the option with the largest probability, then he is not off the hook.

Given that $P(\text{Responsible for } Choice = 1 | Attention = 0) = 1$, and $P(\text{Responsible for } Choice = 1 | Attention = 1) < 1$, he is not off the hook.

Obvious threshold probability

(Elzein, 2013) proposes similar solution: "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."

Improved robustness

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 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.

Conclusion: **Moderate PAP** is not falsified by Tax Evasion.

Healthy sceptic

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, Pereboom states:

> *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.*

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 at least .000001. Both alternatives are entirely on a par, and hence neither is robust compared to the other.

Example: 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.

Example: Trolley

Clearly 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.

This remains true even if 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.

Again this is very similar to (Elzein,2013): "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."

Outline

Argument against Pereboom

Ginet (2002): different argument against *Tax Evasion*, namely "the timing objection".

Based on the distinction between an agent being responsible for some act A *at time* T and the agent being responsible for A *simpliciter*.

Argument: Joe may well be responsible for $Choice = 1$ at some particular time T, without being responsible for $Choice = 1$ simpliciter. All we need is a robust alternative possibility for $Choice = 1$ at time T.

Joe could have chosen a slightly different time, and then $Choice = 1$ at time $T + 1$.

Counterargument

Are our responsibility judgments really that "fragile"? Let's look at causation, again.

Note that we're back to counterfactual dependence as being necessary: responsibility for $Choice = 1$ at time T implies that $Choice = 1$ at T is counterfactually dependent on the agent's control.

Same argument existed to save necessity of counterfactual dependence for causation: our judgements are sensitive to the precise circumstances, temporally and otherwise, under which the event occurs. Lewis (1986): yes, sometimes our causal judgments are indeed about a "fragile" event.

But not *all* our judgements are fragile!

Example: Late Preemption

Example 4

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.

Suzy's throw is a cause, and yet no counterfactual dependence.

But if we use fragility, we can rescue counterfactual dependence.

Hall and Paul (2013) provide a comprehensive discussion of this proposal, pointing out the severe problems that it faces.

The most obvious one: Billy's throw also becomes a cause.

Back to responsibility

Assume that the bottle contains some irreplaceable life-saving medicine for the terminally ill Jones. Assume that both Suzy and Billy fulfill all non-causal conditions for responsibility, whatever they may be.

So causation is the only possible difference between them. But by using the timing objection, there is no causal difference.

Therefore, Billy is also responsible for Jones' death. That strikes me as absurd.

(Note: Of course Billy may well be responsible for *intending* to kill Jones, and this in itself is enough to consider his action immoral. But that doesn't make him responsible for the outcome itself.)

Pereboom and the timing objection

Pereboom (2012) also responded, but Palmer (2013) offered strong criticism.

Why can't Pereboom use my reply? Because he also invokes a time-indexed causal judgment.

Conclusion

- Strong PAP: requires non-question-begging definition of derivative vs non-derivative
- Weak PAP: consistent with all Frankfurt cases (Frankfurt, Fischer, Mele & Robb, Stump, Pereboom,...)
- Moderate PAP: also consistent with all Frankfurt cases, if we use a sensible definition of **robustness**
- General: Frankfurt debate needs explicit Causal Models

Future Work

Formal definition of Responsibility:

- Causal relation between acts and outcomes
- Epistemic state of agent
- Production vs Causation
- Causation and Responsibility: two sides of the same coin?