Formalizing the Causal Conditions for Moral Responsibility

Sander Beckers

Department of Philosophy and Religious Studies
Utrecht University

REINS conference, February 12, 2019

Outline

Moral Responsibility

Responsibility for *outcomes*, grounded in *choices* made by a *single agent*.

Related notions: accountability, blameworthiness/praiseworthiness.

Focus lies on *formal* analysis of *causal* conditions, not on exhaustive philosophical analysis.

Concretely: no discussion of free will, agency, intentionality, norms.

Simplifying Assumptions

- Ignore epistemic responsibility. (Eg: doctor)
- Only one morally relevant outcome.
  - So *intending* and *foreseeing* are equivalent. (Eg: Trolley)
  - All choices are *eligible*. (Eg: Bank)
- Ignore description-relative issues:
  - Everyone describes events in same unique manner. (Eg: Santa Claus)
  - Everyone agrees on which events are relevant. (Eg: Queen)

Guiding Meta-definition

An agent is responsible for O iff the agent *tried* to *produce* O and was *successful* in doing so.

- *successful*: the agent did in fact produce O.
- *produce*: the objective causal condition (contribute to, bear authorship of, cause,...).
- *tried*: involves making a choice based on expectations.

Informal Definition of BvH

Braham & van Hees: An Anatomy of Moral Responsibility, 2012.

Definition (**Responsibility**)

Given that the outcome O occurred, an agent is responsible for O if some event C occurred such that the following conditions hold:

- **(Agency Condition)** The agent autonomously chose C.
- **(Causal Condition)** C NESS-caused O.
- **(Avoidance Condition)** The agent believed that there exists some "reasonable" C' so that C' was less likely to NESS-cause O.

Problem

They make this formally precise using game theory. But game theory is unable to express causal relations!

BvH reply:

1. We agree, **Causal Condition** should be actual causation, not NESS-causation.
2. Actual causation = Halpern & Pearl (2005).
3. But... for the examples discussed "game-theoretic" NESS-causation is fine.

My reply

They're wrong:

1. **Causal Condition** should be NESS-causation, but properly formalized.
2. Halpern & Pearl (2005) is not a good definition of actual causation. (See Beckers & Vennekens (2017, 2018).)
3. Game theoretic NESS-causation is unable to distinguish between *logical sufficiency* and *causal sufficiency*.
   - Frankfurt case fails.
   - Very simple cases, like Late Preemption, fail.

Structural equations modelling

Introduced by Pearl (2000) "Causality: Models, reasoning and inference".

A *causal model* is a tuple $M = ((\mathcal{U}, \mathcal{V}, \mathcal{D}, \mathcal{R}), \mathcal{F})$:

- $\mathcal{U}$: set of exogenous variables
- $\mathcal{V}$: set of endogenous variables
- $\mathcal{D} \subseteq \mathcal{V}$: set of agent variables
- $\mathcal{R}$: function that determines the possible values for every variable $Y \in \mathcal{U} \cup \mathcal{V}$
- $\mathcal{F}$: set of structural equations (one for each $X \in \mathcal{V}$):
  - E.g., $X = Y \wedge Z$, $A = f(B, C, D)$.

Syntax

- primitive events $X = x$
- $[\vec{X} \leftarrow \vec{x}]\varphi$ ("after setting $\vec{X}$ to $\vec{x}$, φ holds")
- φ can be any propositional combination of primitive events

Semantics

Intuition:

- Structural model represents *counterfactual* relations between variables, that are the result of *interventions*: given that $\vec{X} = \vec{x}$, what happens if we intervene and set $\vec{X} = \vec{x}^*$?
- Interventions are non-backtracking: changes propagate from right-side of equation to left-side, but not vice versa.

Let $\vec{u}$ be a context, i.e., a setting of the exogenous variables:

- $(M, \vec{u}) \models \mathcal{V} = \vec{v}$ if $\mathcal{V} = \vec{v}$ is unique solution to equations given $\vec{u}$.
- $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}]\varphi$ if $(M_{\vec{X} \leftarrow \vec{x}}, \vec{u}) \models \varphi$.
- $M_{\vec{X} \leftarrow \vec{x}}$ is the causal model after setting $\vec{X}$ to $\vec{x}$:
  - replace the original equations for the variables in $\vec{X}$ by $\vec{X} = \vec{x}$.

Probabilistic Causal Model: $M = (\mathcal{S}, \mathcal{F}, \Pr)$ is just a causal model together with a probability Pr on contexts.

Agent variables

$\mathcal{D} \subseteq \mathcal{V}$ are the *agent variables*: under "direct control" of the agent.

We simply accept that in some way or other the agent *directly chooses* a $d_i \in \mathcal{R}(D_i)$ for every $D_i \in \mathcal{D}$ w.r.t. M.

- I.e., an actual story is given by a causal setting of the form $(M_{\mathcal{D} \leftarrow \vec{d}}, \vec{u})$.

Responsibility setting $Resp = \{(M, \vec{u}); (M_A, \Pr), [\mathcal{D} \leftarrow \vec{d}]\}$:

- An objective causal setting $(M, \vec{u})$.
- The agent's probabilistic causal model $(M_A, \Pr)$, where M and M_A have the same signature $(\mathcal{U}, \mathcal{V}, \mathcal{D}, \mathcal{R})$.
- The agent's choices $[\mathcal{D} \leftarrow \vec{d}]$.

Agent uncertainty

Probability Pr on $\mathcal{R}(\mathcal{U})$ induces probability on $\mathcal{R}(\mathcal{V})$:

$$\Pr(\vec{v}) = \Pr(\{\vec{u} : (M_A, \vec{u}) \models \mathcal{V} = \vec{v}\}).$$

Each intervention $\vec{X} \leftarrow \vec{x}$ also induces a probability on $\mathcal{R}(\mathcal{V})$:

$$\Pr^{\vec{X} \leftarrow \vec{x}}(\vec{v}) = \Pr(\{\vec{u} : (M_A, \vec{u}) \models [\vec{X} \leftarrow \vec{x}]\mathcal{V} = \vec{v}\}).$$

We are interested in $\Pr^{\mathcal{D} \leftarrow \vec{d}}$.

Formal Definition

Definition (**Responsibility**)

Given a responsibility setting *Resp* such that $(M_{\mathcal{D} \leftarrow \vec{d}}, \vec{u}) \models O = o$, the agent is responsible for the outcome $O = o$ if the agent chose $O = o$ or the following conditions hold:

- **(Agency Condition)** The agent chose $\mathcal{D} = \vec{d}$.
- **(Causal Condition)** $\mathcal{D} = \vec{d}$ contains a NESS-cause of $O = o$.
- **(Doxastic Condition)**

$$\exists \vec{d}' \neq \vec{d} \in \mathcal{R}(\mathcal{D}) : \Pr^{\mathcal{D} \leftarrow \vec{d}}(\mathcal{D} = \vec{d} \text{ contains a NESS-cause of } O = o)$$

$$>$$

$$\Pr^{\mathcal{D} \leftarrow \vec{d}'}(\mathcal{D} = \vec{d}' \text{ contains a NESS-cause of } O = o)$$

Richard Wright (1985, 1988, 2011).

> *According to the NESS account as initially elaborated, a condition c was a cause of a consequence e if and only if it was necessary for the sufficiency of a set of existing antecedent conditions that was sufficient for the occurrence of e. The required sense of sufficiency, which ... I call 'causal sufficiency' to distinguish it from mere lawful strong sufficiency, is the instantiation of all the conditions in the antecedent ('if' part) of a causal law, the consequent ('then' part) of which is instantiated by the consequence at issue.*

Step 1:

Definition (**Sufficient**)

We say that $\vec{X} = \vec{x}$ is *sufficient* for $E = e$ w.r.t. M if $f_E(\vec{x}) = e$. (The equation for E is: $E = f_E(\mathcal{V} - \{E\})$.)

Step 2

> *According to the NESS account as initially elaborated, a condition c was a cause of a consequence e if and only if it was necessary for the sufficiency of a set of existing antecedent conditions that was sufficient for the occurrence of e.*

Definition (**Contribute**)

Given $(M, \vec{u}) \models C = c \wedge E = e$, we say that $C = c$ *contributes to* $E = e$ if there exists a $\vec{X} = \vec{x}$ containing $C = c$, such that $(M, \vec{u}) \models \vec{X} = \vec{x}$ and $\vec{X} = \vec{x}$ is sufficient for $E = e$, but $\vec{X} = \vec{x} - \{C = c\}$ is not. We call $\vec{X} = \vec{x}$ a *witness* for $C = c$ w.r.t. $E = e$.

Step 3

2nd version of the NESS-definition:

> *or (as is more often the case) if c is connected to e through a sequence of such instantiations of causal laws.*

Definition (**NESS-causation**)

Given $(M, \vec{u}) \models C = c \wedge E = e$, we say that $C = c$ *NESS-causes* $E = e$ if there exists a sequence $C = c, \ldots, V_i = v_i, \ldots E = e$ so that each $V_j = v_j$ contributes to $V_{j+1} = v_{j+1}$.

Formal Definition

Definition (**Responsibility**)

Given a responsibility setting *Resp* such that $(M_{\mathcal{D} \leftarrow \vec{d}}, \vec{u}) \models O = o$, the agent is responsible for the outcome $O = o$ if the agent chose $O = o$ or the following conditions hold:

- **(Agency Condition)** The agent chose $\mathcal{D} = \vec{d}$.
- **(Causal Condition)** $\mathcal{D} = \vec{d}$ contains a NESS-cause of $O = o$.
- **(Doxastic Condition)**

$$\exists \vec{d}' \neq \vec{d} \in \mathcal{R}(\mathcal{D}) : \Pr^{\mathcal{D} \leftarrow \vec{d}}(\mathcal{D} = \vec{d} \text{ contains a NESS-cause of } O = o)$$

$$>$$

$$\Pr^{\mathcal{D} \leftarrow \vec{d}'}(\mathcal{D} = \vec{d}' \text{ contains a NESS-cause of } O = o)$$

Intermezzo

Simplified version of (Beckers and Vennekens, 2017):

Definition (**Actual Causation**)

Given a causal setting $(M, \vec{u})$ such that $(M, \vec{u}) \models C = c \wedge O = o$, we say that $C = c$ *caused* $O = o$ if

- $C = c$ NESS-caused $O = o$; and
- $\exists c' \neq c \in \mathcal{R}(C) : \Pr^{C \leftarrow c'}(C = c' \text{ NESS-causes } O = o) < 1$.

Simple example

BvH:

Example (Two Assassins)

Two assassins, in place as snipers, shoot and kill Victim, with each of the bullets fatally piercing Victim's heart at exactly the same moment.

Intuition: each of them is responsible, *even if they believe the other assassin will certainly be accurate*.

(Causal Condition)

Causal model:

$$Death = Assassin_1 \vee Assassin_2.$$

Minimally sufficient sets for $Death = 1$:

$$\{Assassin_1 = 1\}, \text{ and } \{Assassin_2 = 1\}.$$

So both are NESS-causes.

(Doxastic Condition)

$$\Pr^{Assassin_1 \leftarrow 1}(Assassin_1 = 1 \text{ contains a NESS-cause of } Death = 1) = 1$$

$$>$$

$$\Pr^{Assassin_1 \leftarrow 0}(Assassin_1 = 0 \text{ contains a NESS-cause of } Death = 1) = 0$$

Remember my claims

1. **Causal Condition** should be NESS-causation, but properly formalized.
2. Halpern & Pearl (2005) is not a good definition of actual causation.
3. Game theoretic NESS-causation is unable to distinguish between *logical sufficiency* and *causal sufficiency*.
   - Frankfurt case fails.
   - Very simple cases, like Late Preemption, fail.

Late Preemption

Example (Two Assassins 2)

Two assassins, in place as snipers, shoot Victim, who dies. $Assassin_1$ was slightly faster, so that only his bullet fatally pierced Victim's heart. $Assassin_2$'s bullet arrived when Victim was already dead.

Intuition: only $Assassin_1$ is responsible for Victim's death (but $Assassin_2$ is blameworthy for *attempting* to kill Victim!).

BvH would get that $Assassin_2 = 1$ is also a NESS-cause of $Death = 1$.

NESS-causation

$$Death = Bullet_1 \vee Bullet_2$$
$$Bullet_1 = Assassin_1$$
$$Bullet_2 = Assassin_2 \wedge \neg Bullet_1$$

$Assassin_1 = 1$ is sufficient for $Bullet_1 = 1$, which is sufficient for $Death = 1$, so $Assassin_1 = 1$ is a NESS-cause of $Death = 1$.

$Assassin_2 = 1$ is not sufficient for $Bullet_2 = 1$, and hence not a NESS-cause.

1. **Causal Condition** should be NESS-causation, but properly formalized.
2. Halpern & Pearl (2005) is not a good definition of actual causation.
3. Game theoretic NESS-causation is unable to distinguish between *logical sufficiency* and *causal sufficiency*.
   - Frankfurt case fails.
   - Very simple cases, like Late Preemption, fail.

Example (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 rare 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.

Agent variable *Inject*, context such that $Inject = 1$.

Intuition: Jones is responsible.

Doxastic Condition is easy, because M_A is:

$$Death = Reaction$$
$$Reaction = Inject$$

$$\Pr^{Injection \leftarrow 1}(Injection = 1 \text{ contains a NESS-cause of } Death = 1) = 1$$
$$>$$
$$\Pr^{Injection \leftarrow 0}(Injection = 0 \text{ contains a NESS-cause of } Death = 1) = 0$$

Causal Condition: did $Inject = 1$ cause $Death = 1$?

Objective causal model:

$$Death = Reaction \vee \neg Inject$$
$$Reaction = Inject$$

All counterfactual definitions say "No!", NESS-causation says "Yes!":

(Note that all definitions agree that $Inject = 1$ causes $Reaction = 1$, and $Reaction = 1$ causes $Death = 1$.)

No!

What if we just say that Jones is not responsible? Does that save the counterfactual definitions of actual causation?

Frankfurt case

Example (Frankfurt Version of Injection)

(cont.) ... Imagine that unbeknownst to Jones, Smith is standing behind the curtains, watching Jones' every move. If it were to become clear that Jones would not inject the medicine, Smith would shoot Patient.

$$Death = Reaction \vee Smith.$$
$$Reaction = Inject.$$
$$Smith = \neg Inject.$$

$$Death = Reaction \vee \neg Inject$$
$$Reaction = Inject$$

Frankfurt case

Halpern & Pearl (and many others, but not me): $\mathit{Inject} = 1$ caused $\mathit{Death} = 1$.

So just adding some intermediate event is enough to flip a causation judgment!

So just adding some intermediate event is enough to flip a responsibility judgment!

NESS-causation does not do this. It's a NESS-cause in both examples. (In fact, counterfactuals don't matter.)

1. **Causal Condition** should be NESS-causation, but properly formalized.
2. Halpern & Pearl (2005) is not a good definition of actual causation.
3. Game theoretic NESS-causation is unable to distinguish between *logical sufficiency* and *causal sufficiency*.
   - Frankfurt case fails.
   - Very simple cases, like Late Preemption, fail.

NESS-causation according to BvH:

$$\begin{aligned} \mathit{Death} &= \mathit{Reaction} \vee \mathit{Smith}. \\ \mathit{Reaction} &= \mathit{Inject}. \\ \mathit{Smith} &= \neg \mathit{Inject}. \end{aligned}$$

Note that for all $\vec{u} \in \mathcal{R}(\mathcal{U})$, we have $(M, \vec{u}) \models \mathit{Death} = 1$.

Therefore, $\emptyset$ is logically sufficient for $\mathit{Death} = 1$ w.r.t. M.

Conclusion: Jones can never be responsible, regardless of his beliefs.

Addendum to my criticism

p. 626, BvH:

> *However, by way of a preliminary response to this objection, there is nothing in Frankfurt's discussion of the case that gives us reason not to suppose ... that the game is one of complete information.*

Throughout the entire literature, it is considered crucial that the game is *not* one of complete information!

Conclusion

We considered a simplified setting for moral responsibility, to focus on the causal conditions.

Formulated a general meta-definition of moral responsibility to guide us.

Used the skeleton of Braham & van Hees's definition, but moved it from game theory to causal models.

Formalized NESS-causation on our way to formalizing responsibility.

Applied formal definition of moral responsibility to examples.

Future Work

Drop the simplifying assumptions:

- Add epistemic responsibility: focus on "reasonable" agent models.
- Combine and weigh morally relevant outcomes.
  - So *intending* and *foreseeing* can come apart.
- Add description-relative issues:
  - Focus on properties that hold in all "appropriate" causal models, to reduce model-relativity.

The roots of responsibility: is actual causation just a spin-off from responsibility?