Moral Responsibility for AI Systems

Forthcoming at NeurIPS 2023

Sander Beckers

Institute for Logic, Language and Computation, University of Amsterdam
srekcebrednas@gmail.com
sanderbeckers.com

Amsterdam Causality Meeting, 09/10/2023

Outline

Outline

Moral Responsibility

Responsibility is an overloaded and vague concept:

- Captain on a ship, CEO of a company
- Epistemic component of responsibility: uneducated surgeon
- Derivative responsibility: drunk driving
- Accountability vs attributability vs causal responsibility

Responsibility for a *single outcome*, grounded in a *single choice* made by a *single* and *artificial* agent.

Necessary but not sufficient for: blame or praise.

Moral Responsibility

Contributions:

- Formalize Causal Condition and Epistemic Condition
- Compare to two competing accounts (BvH and HK)
- Both a qualitative and a quantitative definition

Purpose:

- Definition can be used by regulator to evaluate AI systems
- Definition can be used *by an AI system itself* to make responsible choices
- Definition can be used by regulatory AI to evaluate *other* AI systems
- Contributes to philosophical debate on responsibility more generally

Caveat: requires (partial) knowledge of a causal model

Guiding Meta-definition

An agent who performs $A = a$ is responsible for outcome $O = o$ if:

1. The agent had control over $A = a$.
2. $A = a$ causes $O = o$.
3. The agent believes there exists a' so that by performing $A = a'$ they would have avoided being responsible for $O = o$.

(2) is the **Causal Condition**: $A = a$ is an actual cause of $O = o$.

(3) is the **Epistemic Condition**

Informal Definition of BvH

Braham & van Hees (2012) An Anatomy of Moral Responsibility, *Mind*.

Formalism: game-theory

Definition (**BvH Responsibility**)

- **(Causal Condition)** $A = a$ directly NESS-causes $O = o$.
- **(Epistemic Condition)** $A = a$ does not minimize probability of causing $O = o$.

Informal Definition of HK

Halpern & Kleiman-Weiner (2018) Towards Formal Definitions of Blameworthiness, Intention, and Moral Responsibility, *AAAI 18*.

Formalism: causal models + utilities

Definition (**HK Responsibility**)

- **(Causal Condition)** $A = a$ HP-causes $O = o$.
- **(Epistemic Condition)** $A = a$ does not minimize probability of $O = o$.

Choices to be made

- Which formalism? (Game-theory vs causal models)
- Which definition of causation?
  - Necessary Element of a Sufficient Set (NESS)
  - Halpern & Pearl (HP)
  - Counterfactual NESS (CNESS)
- Minimization: what to minimize?
  - **(Outcome)** (HK):

$$Pr(O = o|do(A = a)) \leq Pr(O = o|do(A = a'))$$

  - **(Causality** (BvH):

$$Pr(A = a \text{ causes } O = o) \leq Pr(A = a' \text{ causes } O = o)$$

  - **Combination of both**

My Proposal

Formalism: causal models

Definition (**Responsibility**)

- **(Causal Condition)** $A = a$ CNESS-causes $O = o$.
- **(Epistemic Condition)**
  - $A = a$ does not minimize probability of $O = o$, or
  - $A = a$ only minimizes the probability of $O = o$.

Further step: degree of responsibility

Arguments

1. NESS (let alone HP) cannot be captured using game-theory
   - direct NESS vs indirect NESS
2. CNESS $>$ NESS $>$ direct NESS, and CNESS $>$ HP
   - My other work, but also some examples
3. Preventing outcome is priority, but it's not enough
   - Example where both conditions conflict

A word about causation

- Necessary Element of a Sufficient Set (NESS)
  - Richard Wright, John Mackie (INUS), legal philosophy, regularity approach.
- Halpern & Pearl (HP)
  - Causal models, counterfactual approach, AI, 2001-2005-2016.
- Counterfactual NESS (CNESS)
  - Causal models, counterfactual *and* regularity approach, based on Wright
  - Beckers (2021) The Counterfactual NESS Definition, *AAAI 2021*.
  - Simplification of Beckers (2021) Causal Sufficiency and Actual Causation, *Journal of Philosophical Logic*.

Outline

Informal BvH Definition

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.

(Causal Condition):

- $A_1 = 1$ is sufficient for $V = 1$
- $\emptyset$ is not sufficient for $V = 1$
- So $A_1 = 1$ NESS-causes $Death = 1$.

(Epistemic Condition):

- $Pr(A_1 = 1 \text{ NESS-causes } V = 1) = 1$
- $Pr(A_1 = 0 \text{ NESS-causes } V = 1) = 0$.
- So $A_1 = 1$ fails to minimize.

Likewise for $A_2 = 1$.

Informal BvH Definition

Example (**Late Preemption)**

Assassin$_1$ is slightly faster, so that his bullet kills Victim, who collapses and thereby dodges *Assassin*$_2$'s bullet.

$A_2 = 1$ *does not* cause $V = 1$!

(Causal Condition):

- $A_2 = 1$ is sufficient for $V = 1$
- $\emptyset$ is not sufficient for $V = 1$
- So $A_2 = 1$ NESS-causes $V = 1$....

Causal Models

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

- $\mathcal{U}$: set of exogenous variables
- $\mathcal{V}$: set of endogenous 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 $X \in \mathcal{V}$):

Late Preemption:

- $V = BH_1 \vee BH_2$
- $BH_1 = A_1$
- $BH_2 = A_2 \wedge \neg BH_1$

Direct NESS

Definition (**Sufficiency**)

We say that $\vec{X} = \vec{x}$ is *sufficient* for $Y = y$ w.r.t. $(M, \vec{u})$ if for all $\vec{z}$ we have that $Y_{\vec{x},\vec{z}}(\vec{u}) = y$.

In our example:
$BH_1 = A_1$: therefore $A_1 = 1$ is sufficient for $BH_1 = 1$.

$BH_2 = A_2 \wedge \neg BH_1$: therefore $A_2 = 1$ is **not** sufficient for $BH_2 = 1$.

Also: $A_1 = 1$ is **not** sufficient for $V = 1$.

Direct NESS

- the candidate cause and the effect actually occurred;
- the candidate cause is a member of a sufficient set;
- and it is necessary for the set to be sufficient.

Definition (**Direct NESS**)

$X = x$ *directly NESS-causes* $Y = y$ w.r.t. $(M, \vec{u})$ if there exists a $\vec{W} = \vec{w}$ so that the following conditions hold:

DN1. $(M, \vec{u}) \models X = x \wedge \vec{W} = \vec{w} \wedge Y = y$.

DN2. $X = x \wedge \vec{W} = \vec{w}$ is sufficient for $Y = y$ w.r.t. $(M, \vec{u})$.

DN3. $\vec{W} = \vec{w}$ is not sufficient for $Y = y$ w.r.t. $(M, \vec{u})$.

from Direct NESS to NESS

Late Preemption:

- $V = BH_1 \vee BH_2$
- $BH_1 = A_1$
- $BH_2 = A_2 \wedge \neg BH_1$

Context $\vec{u}$: $A_1 = 1$ and $A_2 = 1$

$A_1 = 1$ directly NESS-causes $BH_1 = 1$

$BH_1 = 1$ directly NESS-causes $V = 1$

NESS-causation: transitive closure of direct NESS-causation *along a path*

So $A_1 = 1$ NESS-causes $V = 1$ along $\{A_1, BH_1, V\}$.

Example (**One Assassin**)

Assassin$_1$ does not shoot, so that Victim is killed by *Assassin*$_2$'s shot. As before, *Assassin*$_1$ is the faster shooter, so had he shot, then it would have been his bullet that got to Victim first.

Assassin$_1$ is obviously not responsible for Victim's death.

(Causal Condition):

- $A_1 = 0$ is sufficient for $BH_1 = 0$.
- $BH_1 = 0 \wedge A_2 = 1$ is sufficient for $BH_2 = 1$,
- whereas $A_2 = 1$ is not.
- $BH_1 = 0 \wedge A_2 = 1$ is sufficient for $BH_2 = 1$.
- $BH_2 = 1$ is sufficient for $Death = 1$.
- So $A_1 = 0$ NESS-causes $V = 1$ along the path $\{A_1, BH_1, BH_2, V\}$.

(Epistemic Condition): flare gun to warn *Victim*

Outline

Counterfactual NESS

The Counterfactual NESS Definition of Causation, AAAI 2021

Definition (**CNESS-causation)**

$C = c$ *CNESS-causes* $E = e$ if

- $C = c$ NESS-causes $E = e$ along some path p and
- there exists a c' such that $C = c'$ would not have NESS-caused $E = e$ along any subpath p' of p.

Counterfactual NESS

One Assassin Example:

- $A_1 = 0$ NESS-causes $V = 1$ along the path $\{A_1, BH_1, BH_2, V\}$.
- $A_1 = 1$ NESS-causes $V = 1$ along the path $\{A_1, BH_1, V\}$.
- $\{A_1, BH_1, V\} \subseteq \{A_1, BH_1, BH_2, V\}$.
- So $A_1 = 0$ is not a CNESS-cause of $V = 1$.

Against HP-causation

Beckers, S. (2021) The Counterfactual NESS Definition of Causation, AAAI.

Beckers, S. (2021) Causal Sufficiency and Actual Causation, Journal of Philosophical Logic.

Against HP-causation

Example (**Loader**)

"Suppose that a prisoner dies either if A loads B's gun and B shoots, or if C loads and shoots his gun. A loads B's gun, B does not shoot, but C does load and shoot his gun, so that the prisoner dies. We would not want to say that $A = 1$ is a cause of $D = 1$, given that B did not shoot (i.e., given that $B = 0$)." (HP 2005)

- $D = (A = 1 \wedge B = 1) \vee C = 1$

$A = 1$ does not HP-cause $D = 1$

Example (**Loader 2**)

C only fired his gun because B did not shoot ($C = \neg B$).

$A = 1$ HP-causes $D = 1$

Outline

Two Lessons

1. Preventing the outcome matters more than preventing causing the outcome

2. Yet preventing causing the outcome does matter

Example (**Bomb**)

A bomb (B) is connected to three detonators (D_1, D_2, and D_3) by two switches (S_1 and S_2). D_1 is functional if only S_1 is on, D_2 is functional if only S_2 is on, and D_3 is functional whenever S_1 is on.

- $B = D_1 \vee D_2 \vee D_3$
- $D_1 = S_1 \wedge \neg S_2$
- $D_2 = S_2 \wedge \neg S1$
- $D_3 = S_1$
- $Pr(S_1 = 1) = 0.6$

$Assassin_2$ decides to turn on S_2, thereby guaranteeing that the bomb will explode. $Assassin_1$ decides not to turn on S_1, so that the bomb explodes only due to the functioning of D_2.

Causal Condition: $S_2 = 1$ causes $B = 1$

Intuition: $Assassin_2$ is responsible for $B = 1$

Preventing Outcome (**HK**):

$$P(B = 1|do(S_2 = 1)) = 1$$

$$>$$

$$P(B = 1|do(S_2 = 0)) = 0.6$$

Preventing Causation (**BvH**):

$$P(S_2 = 1 \text{ causes } B = 1) = 0.4$$

$$<$$

$$P(S_2 = 0 \text{ causes } B = 1) = 0.6$$

Example (**Two Assassins**)

$Pr(A_2 = 1) = 1$, so $Pr(V = 1) = 1$

So $Assassin_1$ minimizes probability of outcome.

But he is still responsible!

Moral of the story:

- Priority: try to prevent outcome
- If successful: try to prevent causing outcome

Definition (**Responsibility)**

An agent who performs $A = a$ is responsible for outcome $O = o$ w.r.t. a responsibility setting $M, \vec{u}, \mathcal{E})$ if:

(Causal Condition) $A = a$ CNESS-causes $O = o$ w.r.t. $(M, \vec{u})$.

(Epistemic Condition)
There exists $a' \in \mathcal{R}(A)$ so that one of the following holds:

- $\Pr(O = o | do(A = a)) > \Pr(O = o | do(A = a'))$
-

$$\Pr(O = o | do(A = a)) = \Pr(O = o | do(A = a'))$$

and

$\Pr(A = a$ CNESS-causes $O = o > \Pr(A = a'$ CNESS-causes $O = o)$.

Conclusion

Choices:

- Formalism: causal models
- Actual Causation = Counterfactual NESS
- Epistemic Condition: give priority to Actuality Condition, but do not forget Causal Condition.

Future work:

- Multiple outcomes/agents/actions
- Extend to blame and praise
- Incorporate Harm:
  - Beckers, S., Chockler, H., and Halpern, J.Y. (2022). A Causal Analysis of Harm, *NeurIPS 2022*.
  - Beckers, S., Chockler, H., and Halpern, J.Y. (2023). Quantifying Harm, *IJCAI 2023*.