Moral Responsibility

Dean McHugh

Institute of Logic, Language and Computation
University of Amsterdam

Causal Inference Lab
Master of Logic project
8 January 2020

Institute for Logic, Language and Computation

NWO

University of Amsterdam

Outline

1. Responsibility and causation
   - Is causation necessary for moral responsibility?
   - Causation beyond one-shot games
   - Moral responsibility in practice versus in theory

2. Overdetermined outcomes
   - Responses to overdetermination
   - A breezy introduction to game theory
   - Case study: Norway's oil industry

Outline

1. Responsibility and causation
   - Is causation necessary for moral responsibility?
   - Causation beyond one-shot games
   - Moral responsibility in practice versus in theory

2. Overdetermined outcomes
   - Responses to overdetermination
   - A breezy introduction to game theory
   - Case study: Norway's oil industry

Is causation necessary for moral responsibility?

Braham and van Hees (2012): Yes.

Testing Braham and van Hees' definition

Example (Army chain of command)

- If there is a conflict between officers, the soldiers obey the most senior officer
- Both the senior and junior officer order: CHARGE!
- Their attack is later found to be a war crime

Senior	Junior	Charge
Yes	Yes	✓
Yes	Silent	✓
Yes	No	✓
Silent	Yes	✓
Silent	Silent	✗
Silent	No	✗
No	Yes	✗
No	Silent	✗
No	No	✗

1. Did the junior officer *causally contribute* to the soldiers' charging?
2. Is the junior officer (in part) *morally responsible* for the outcome?
3. What do Braham and van Hees (2012) predict?
    - The junior officer did not cause the soldiers to charge
    - ∴ The junior officer is not morally responsible

A methodological note

A worry

To get clear answers in real-life cases using logical methods, we have to idealise the scenario, and make many unrealistic assumptions

We might assume, for instance,

- That the junior officer acted independently from the senior officer
- That the junior officer satisfies an agency condition

Are these assumptions realistic?

A response:

- Idealising assumptions are a temporary strategy
- Even after simplification, the idealised cases are still difficult
  - Once we have clear answers to the idealised cases, we can try to consider more realistic and complicated cases

Extra components of responsibility

> *the problem of assessing degrees of responsibility for an outcome is not only a matter of assessing the extent of each individual's causal contribution to that outcome but also involves an assessment and integration of various other dimensions such as degrees of initiative, degrees of authority, the gains from the activities involved, and perhaps most difficult of all, the degree of voluntariness.*
>
> — Braham and van Hees (2009, p. 341)

Outline

1. Responsibility and causation
   - Is causation necessary for moral responsibility?
   - Causation beyond one-shot games
   - Moral responsibility in practice versus in theory

2. Overdetermined outcomes
   - Responses to overdetermination
   - A breezy introduction to game theory
   - Case study: Norway's oil industry

Causation beyond the Braham & van Hees definition

Braham and van Hees define causation over one-shot games
But some causal information does not fit in this framework

Example (Two ways to count votes)

- A committee with two members: Chair and CEO
- New company policies require both to vote Yes to pass
- Two secretaries, each with their own way to declare the result:

Both — Opens both envelopes and simultaneously checks each vote. The motion passes just in case both say 'yes'.

CEO first — Opens the CEO's envelope first: if it says 'yes' they then open the Chair's envelope and see whether it says 'yes' too, in which case the motion passes. If the CEO votes 'no' the secretary declares the motion failed without ever looking inside the Chair's envelope.

Example (Counting votes)

- New company policies require both to vote Yes to pass
- Suppose each committee member does not know how the other will vote
  - Further (if needed), suppose that each is unaware of which counting rule is in operation
- Suppose the CEO first method is used
- Scenario:
  - The Chair and CEO both vote No
  - The CEO's vote is read: the motion fails immediately
  - The Chair's envelope is never opened (ever, let's say)
  - The rejection of the proposal led to the company's bankruptcy

1. Did the Chair's vote causally contribute to the outcome?
2. Is the Chair morally responsible for the outcome?

In the *Counting Votes* scenario,

- It seems the Chair's vote did not causally contribute to the outcome
  - How could a vote that was never seen make a causal contribution?
- But it seems the Chair is morally responsible for the outcome
  - Unlike the voting rule (which is an AND-gate, and hence commutative), the counting rule is arbitrary; e.g. it does not change each voter's power over the result of the vote
  - The counting rule might make a causal difference, but it does not make a moral difference

Result

Responsibility without causation

Causation in fact vs. believed causation

A proposal

To be held morally responsible for an outcome (that actually occurred),

- We do not require that one in fact causally contributed to it
- Rather, we require that it is consistent with what one believes that one causally contributed to it

In other words, a necessary condition for moral responsibility is that

- the scenario is indistinguishable (from one's perspective) from a scenario where one actually did causally contribute to the outcome

One way to save the causation–responsibility link

One way to keep that causation is necessary for responsibility:

- We saw two cases where an agent is intuitively morally responsible for an outcome without causing it (junior officer; Chair)
- But perhaps we misinterpreted the scenarios
  - Proposal: keep better track of the *outcomes*
  - Being morally responsible *for an outcome* requires causally contributing *to that outcome*
  - Even when an agent did not causally contribute to an outcome, we can still hold them morally responsible for a *different* outcome; namely, their attempt to cause the outcome
- The junior officer and Chair are not morally responsible for the outcome, but for attempting to cause it
- Just as attempted murder is itself a crime, the junior officer and Chair are blameworthy for attempting to cause the outcome

A reply to moral responsibility for an attempt

The proposal claims that in some situations, an agent is morally responsible for an outcome because

- They attempted to cause it
- And satisfy the remaining conditions for responsibility (such as the agency and avoidance opportunity conditions)

However, there are situations where an agent satisfies these conditions but we do *not* hold them morally responsible...

Attempting to cause without moral responsibility

Example (Counting votes + attempt)

An agent attempts to take the Chair's place on the committee so they can vote No on the proposal and make the motion fail. They do not succeed at taking the Chair's place, and so do not get the chance to vote.
The original Chair and CEO vote No, the CEO first secretary opens the CEO's vote, and the proposal fails (without seeing the Chair's vote).

- Intuitively, the agent who attempted to take the Chair's place is not morally responsible for the motion's failure
- But they did attempt to cause the motion to fail

If one can be responsible for an outcome by attempting to cause it,

- What distinguishes the Chair's role in *Counting votes* from the attempted Chair in *Counting votes + attempt*?
- And in general, under what conditions do we hold an agent morally responsible for attempting to cause a bad outcome?

Reply appealing to beliefs about causation

In the *Counting votes + attempt* scenario,

- The Chair believed that their vote could have causally contributed to the outcome
  - Because the Chair did not know how the CEO would vote (or, let's say, which counting rule was in operation)
- The agent who attempted to replace the Chair did not believe that they could have causally contributed to the outcome
  - Because, e.g., that agent knew they did not make it onto the committee, and so did not even get to vote

Outline

1. Responsibility and causation
   - Is causation necessary for moral responsibility?
   - Causation beyond one-shot games
   - Moral responsibility in practice versus in theory

2. Overdetermined outcomes
   - Responses to overdetermination
   - A breezy introduction to game theory
   - Case study: Norway's oil industry

Moral responsibility in practice versus in theory

The avoidance opportunity condition (AOC)

The agent should have had a reasonable opportunity to have done otherwise.

Braham and van Hees' analysis: There was some alternative course of action available to agent that would have reduced their chance of causally contributing to the outcome.

Braham and van Hees (2012):

- Satisfying the AOC is necessary to be held morally responsible
- Intuitive support: "You can't hold me morally responsible for the outcome: I did everything in my power to try to *prevent* it!"

The reasons we care about moral responsiblity

Example (Making the best of a bad situation)

A patient is in a coma for many years. The family decide to take the patient off of life support, and the patient dies.

According to the analysis of Braham and van Hees (2012), are the family morally responsible for the patient's death?

Agency: ✓ The family were free agents who made their decision consciously

Causation: ✓ The family caused the patient's death

Avoidance opportunity: ✓ The family could have kept the patient on life support

Still, would we want to assert that the family are morally responsible for the patient's death?

Praise and blame, and responsibility

- There is an intuitive connection between responsibility, on the one hand, and praise and blame, on the other
- Indeed, perhaps the whole reason we care about moral responsibility comes from its role in assigning praise and blame
- Compare *Making the best of a bad situation* (previous slide) with making the worst of a good situation (e.g. killing a sentient person)
- Moral responsibility comes apart from blameworthiness here
    - In both cases, the agent is morally responsible for a death
- So what, if not moral responsibility, explains our different reactions to the two cases (making the best of a bad situation *vs.* making the worst of a good situation)
- Perhaps we should look to blameworthiness

Blameworthiness and expected welfare

A Principle

Agents should aim to maximise expected social welfare

Advantages

1. Explains why one is not blameworthy for making the best of a bad situation
2. Explains why moral responsibility concerns one's causal beliefs rather than causation in fact
    - Expected social welfare: includes the agents' beliefs (credence/probabilities) over outcomes

Outline

1. Responsibility and causation
    - Is causation necessary for moral responsibility?
    - Causation beyond one-shot games
    - Moral responsibility in practice versus in theory

2. Overdetermined outcomes
    - Responses to overdetermination
    - A breezy introduction to game theory
    - Case study: Norway's oil industry

Overdetermined outcomes

Scenarios with overdetermined outcomes:

1. Two assassins both kill a victim at the same time
2. A unanimous vote (bit more realistic)

> *in Two Assassins, by not shooting, an assassin would not have prevented the outcome, but would have ensured that he would not be causally effective for it [...]. Similarly, for the committee example, by voting against the proposal an individual committee member would not have prevented the outcome, but would have ensured that he would not be causally effective for it being approved.*
>
> — Braham and van Hees (2012, p. 608)

Responses to responsibility evasion

A response to the “If I don’t do it, someone else will” argument

If you don’t want to be responsible, let someone else do it and let them be held responsible instead.

A problem with this response

- Sometimes one can (or believes one can) do the action better – i.e. with less harm – than others

Example

Norway’s oil industry https://youtu.be/zSjYra7cYqY?t=240

But first...

Outline

1. Responsibility and causation
   - Is causation necessary for moral responsibility?
   - Causation beyond one-shot games
   - Moral responsibility in practice versus in theory
2. Overdetermined outcomes
   - Responses to overdetermination
   - A breezy introduction to game theory
   - Case study: Norway's oil industry

A proof of concept using game theory

Following slides based on those by Ulle Endriss for the course *Game Theory*

Definition (Normal form game)

A *normal form game* is a tuple $G = (N, \boldsymbol{A}, \boldsymbol{u})$ where

- $N = \{1, \ldots, n\}$ is a finite set of players
- $\boldsymbol{A} = A_1 \times \cdots \times A_n$ is a finite set of action profiles $\boldsymbol{a} = (a_1, \ldots, a_n)$, where for each player $i \in N$, A_i is the set of actions available to i
- $\boldsymbol{u} = (u_1, \ldots, u_n)$ is a profile of utility functions; each $u_i : \boldsymbol{A} \to \mathbb{R}$ assigns the payoff player i receives under action profile $\boldsymbol{a}$

- Players choose their actions simultaneously, without knowledge of the other players' choices
- Each player is assumed to be rational and self-interested (a "*homo economicus*"): their sole aim is to maximise their own utility
- The structure of the game $(N, \boldsymbol{A}, \boldsymbol{u})$ is common knowledge

Example (Prisoners’ dilemma)

B A	silent	betray
silent	-1 -1	0 -3
betray	-3 0	-2 -2

Figure: Prisoners’ dilemma payoff matrix

Optimal strategies in game theory

Notation:

- $\boldsymbol{a}_{-i}$ is the action profile $\boldsymbol{a}$ with player *i*'s action removed:

$$\boldsymbol{a}_{-i} := (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$$

- $(a'_i, \boldsymbol{a}_{-i})$ is the action profile that results from replacing a_i, *i*'s action in $\boldsymbol{a}$, with action a'_i:

$$(a'_i, \boldsymbol{a}_{-i}) := (a_1, \ldots, a_{i-1}, a'_i, a_{i+1}, \ldots, a_n)$$

Definition (Best response)

An action $a_i \in A_i$ is a *best response* for player i to action profile $\boldsymbol{a}_i$ iff

$$u_i(a_i, \boldsymbol{a}_{-i}) \geq u_i(a'_i, \boldsymbol{a}_{-i}) \qquad \text{for every } a'_i \in A_i$$

A best response is an action such that, fixing the actions of the other players, i cannot improve by switching.

Nash equilibria and social welfare

Definition (Pure Nash equilibrium)

An action profile $\boldsymbol{a} = (a_1, \ldots, a_n) \in \boldsymbol{A}$ is a *pure Nash equilibrium* iff a_i is a best response to $\boldsymbol{a}_{-i}$ for every player $i \in N$.

Nash equilibria are stable in the sense that, given the actions of the other players, no player wishes to switch.

Definition (Social welfare)

The *social welfare* of action $\boldsymbol{a}$ is the sum of the players' utilities under $\boldsymbol{a}$:

$$sw(\boldsymbol{a}) := \sum_{i \in N} u_i(\boldsymbol{a})$$

An action profile $\boldsymbol{a}$ is a *social optimum* iff it maximizes social welfare:

$$\boldsymbol{a} \in \arg\max_{\boldsymbol{a}' \in \boldsymbol{A}} sw(\boldsymbol{a}')$$

Expected utility and social welfare

- Let $G = (N, \boldsymbol{A}, \boldsymbol{u})$ be a normal form game
- For each agent $i \in N$, let $p_i : \boldsymbol{A}_{-i} \to [0,1]$ be a probability distribution over the other agents' actions (where $\boldsymbol{A}_{-i} := A_1 \times \cdots \times A_{i-1} \times A_{i+1} \times \cdots \times A_n$)

Definition (Expected utility)

The *expected utility* of playing $a_i \in A_i$ for an agent $i \in N$ is

$$eu_i(a_i) := \sum_{\boldsymbol{a}_{-i} \in \boldsymbol{A}_{-i}} p_i(\boldsymbol{a}_{-i}) u_i(a_i, \boldsymbol{a}_{-i})$$

Definition (Expected social welfare)

The *expected social welfare* of playing $a_i \in A_i$ for an agent $i \in N$ is

$$esw_i(a_i) := \sum_{\boldsymbol{a}_{-i} \in \boldsymbol{A}_{-i}} p_i(\boldsymbol{a}_{-i}) sw(a_i, \boldsymbol{a}_{-i})$$

Example (Prisoners' dilemma)

B A	silent	betrays
silent	-1 -1	0 -3
betrays	-3 0	-2 -2

Figure: Prisoners' dilemma payoff matrix

- Which actions are Nash equilibria?
- Which actions are social optima?

A moral (we already knew): The preferences of self-interested actors are often not socially optimal (to say the least!).

Something to keep in mind when applying game theory

Behavioural economics: the *homo economicus* does not exist!

- Real agents systematically deviate from maximizing expected utility

  See e.g. the prospect theory of Tversky and Kahneman (1974)

- Some real agents are not always self-interested
- Sometimes this limitation of game theory is actually a strength: to the extent that actors are purely self-interested

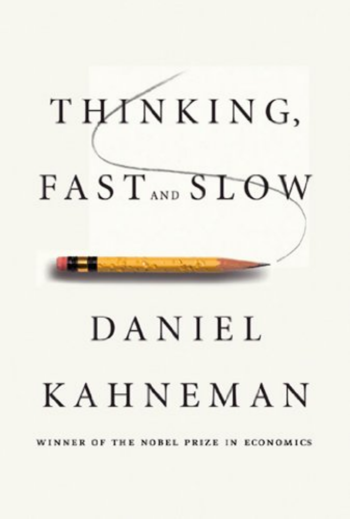


Figure: Kahneman (2011)

I have not seen, in all my life, a single example where a game theorist could give advice, based on the theory, which was more useful than that of the layman.

— Ariel Rubinstein (2012)

Outline

1. Responsibility and causation
   - Is causation necessary for moral responsibility?
   - Causation beyond one-shot games
   - Moral responsibility in practice versus in theory

2. Overdetermined outcomes
   - Responses to overdetermination
   - A breezy introduction to game theory
   - Case study: Norway's oil industry

Norway's oil industry

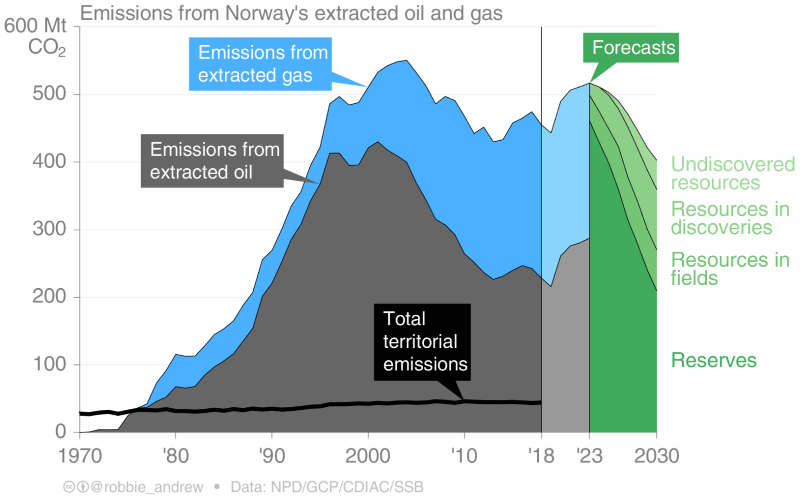
A case study in responsibility

> *Every person has the right to an environment that is conducive to health and to a natural environment whose productivity and diversity are maintained.*
> *Natural resources shall be managed on the basis of comprehensive long-term considerations which will safeguard this right for future generations as well.*
>
> — Constitution of the Kingdom of Norway §112 (translation)

- §112 cited in *The People v Arctic Oil* (16-166674TVI-OTIR/06)
- See also the Judgement of the Oslo District Court[1]

[1] `https://secured-static.greenpeace.org/norway/Global/norway/Arktis/Dokumenter/2018/Judgement%20-%204.%20jan%202017%20-%20Oslo%20District%20Court%20stamped%20version.pdf`

My profit, someone else's pollution



Robbie Andrew, CICERO Center for International Climate Research see `http://folk.uio.no/roberan/t/export_emissions.shtml`

Game theoretic models of responsibility

- Profit from oil sales: $+7$ for the seller
- Harm from oil sales: $\begin{cases} -2 \text{ per country} & \text{if N sells} \\ -4 \text{ per country} & \text{if S sells} \end{cases}$
- N has the chance to sell; if it does not, S will sell instead

Example (Oil sales)

S *N*	sell if *N* ¬sell	¬sell if *N* ¬sell
sell	−2 5	−2 5
¬sell	3 −4	0 0

- Nash equilibrium:
  Norway sells the oil
- Social optimum:
  Norway sells the oil

Increasing the costs

- Profit from oil sales: $+7$ for the seller
- Harm from oil sales: $\begin{cases} -4 \text{ per country} & \text{if N sells} \\ -6 \text{ per country} & \text{if S sells} \end{cases}$

High environmental cost

S N	sell if N ¬sell	¬sell if N ¬sell
sell	-4 3	-4 3
¬sell	1 -6	0 0

- Nash equilibrium: Norway sells the oil
- Social optimum: No one sells the oil

Responsibility in normal form games

- Without further refinement, normal form games do not represent responsibility relations
- Why? One reason (among others): normal form games do not encode causation
- One can seek to add information representing causation to normal form games For example, see Braham and van Hees (2012)

A very simple way to encode causation in games

Specify the *contribution* each player makes to the other players' utilities

i.e. for any players i and j and action profile $\boldsymbol{a}$, specify how much of i's utility under $\boldsymbol{a}$ is attributed to the action j performed in $\boldsymbol{a}$.

Definition (Attribution game)

An *attribution game* is a tuple $(N, \boldsymbol{A}, \boldsymbol{u})$ where

- N is a set of players and $\boldsymbol{A}$ a set of action profiles, as before
- $\boldsymbol{u} = (u_1, \ldots, u_n)$ is a profile of utility functions with *individual attributions*; each $u_i : \boldsymbol{A} \times (N \cup \{*\}) \to \mathbb{R}$ is a function where
  - $u_i(\boldsymbol{a}, j)$ represents the portion of the payoff player i receives under action profile $\boldsymbol{a}$ that is attributed to player j in $\boldsymbol{a}$
  - $u_i(\boldsymbol{a}, *)$ represents the portion of the payoff player i receives under action profile $\boldsymbol{a}$ that is not attributed to any player

Every attribution game $(N, \boldsymbol{A}, \boldsymbol{u})$ induces a normal form game $(N, \boldsymbol{A}, \boldsymbol{u}')$ by summing up the portions of attributed utility:

$$u'_i(\boldsymbol{a}) := u_i(\boldsymbol{a}, 1) + \cdots + u_i(\boldsymbol{a}, n) + u_i(\boldsymbol{a}, *).$$

Responsibility and reparation

For any distinct players i, j, let $\textsc{undo}_i^j$ be an operation on utility profiles that denotes "undoing" j's contribution to i.

$$\begin{aligned} \textsc{undo}_i^j(u_i(\boldsymbol{a})) &= u_i(\boldsymbol{a}) - u_i(\boldsymbol{a}, j) \\ \textsc{undo}_i^j(u_j(\boldsymbol{a})) &= u_j(\boldsymbol{a}) + u_i(\boldsymbol{a}, j) \\ \textsc{undo}_i^j(u_k(\boldsymbol{a})) &= u_k(\boldsymbol{a}) && \text{for any } k \text{ with } i \neq k \neq j \end{aligned}$$

We extend the $\textsc{undo}$ operation to groups of agents. For any collection of players $C \subseteq N$ with $j \notin C$,

$$\begin{aligned} \textsc{undo}_C^j(u_i(\boldsymbol{a})) &= u_i(\boldsymbol{a}) - u_i(\boldsymbol{a}, j) && \text{for any } i \in C \\ \textsc{undo}_C^j(u_j(\boldsymbol{a})) &= u_j(\boldsymbol{a}) + \sum_{i \in C} u_i(\boldsymbol{a}, j) \\ \textsc{undo}_C^j(u_k(\boldsymbol{a})) &= u_k(\boldsymbol{a}) && \text{for any } k \text{ with } j \neq k \notin C \end{aligned}$$

Low environmental cost

S N	sell if N ¬sell	¬sell if N ¬sell
sell	-2 5	-2 5
¬sell	3 -4	0 0

(a) Before REPAYing harms

S N	sell if N ¬sell	¬sell if N ¬sell
sell	0 3	0 3
¬sell	-1 0	0 0

(b) After REPAYing harms

Figure: Applying the REPAY function

After repaying harms

- Social optimum: Norway sells the oil
- Nash equilibria: Norway sells the oil

No change from REPAY!

High environmental cost

S N	sell if N ¬sell	¬sell if N ¬sell
sell	−4 3	−4 3
¬sell	1 −6	0 0

(a) Before REPAYing harms

S N	sell if N ¬sell	¬sell if N ¬sell
sell	0 −1	0 −1
¬sell	−5 0	0 0

(b) After REPAYing harms

Figure: Applying the REPAY function

After repaying harms

- Social optimum: No one sells the oil
- Nash equilibrium: No one sells the oil

REPAY changes the Nash equilibrium

The effects of REPAYing

- Low environmental cost
  - Before REPAY
    - Nash equilibrium: Norway sells the oil
    - Social optimum: Norway sells the oil
  - After REPAY
    - Nash equilibrium: Norway sells the oil
    - Social optimum: Norway sells the oil
- High environmental cost
  - Before REPAY
    - Nash equilibrium: Norway sells the oil
    - Social optimum: No one sells the oil
  - After REPAY
    - Nash equilibrium: No one sells the oil
    - Social optimum: No one sells the oil

The effect of REPAY (loosely)
Aligns Nash equilibria with social optima when the harms outweigh the benefits

References I

Matthew Braham and Martin van Hees. Degrees of causation. *Erkenntnis*, 71(3): 323–344, 2009. doi:10.1007/s10670-009-9184-8.

Matthew Braham and Martin van Hees. An anatomy of moral responsibility. *Mind*, 121 (483):601–634, 7 2012. doi:10.1093/mind/fzs081.

Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

Ariel Rubinstein. The best books on game theory, interview by Sophie Roell, 2012. URL `https://fivebooks.com/best-books/ariel-rubinstein-on-game-theory/`.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974. doi:10.1126/science.185.4157.1124.