Moral Responsibility, Blameworthiness, and Intention: In Search of Formal Definitions

Joe Halpern
Cornell University

Includes joint work with Max Kleiman-Weiner (MIT) and Judea Pearl (UCLA).

The Big Picture

What exactly is *moral responsibility* and *intention*?

- People have been discussing these issues for thousands of years.
- Amazon lists over 50 books in Philosophy, Law, and Psychology with the term "Moral Responsibility" in the title
- There are dozens of other books on intention.
- The Cornell library has shelves full of books these topic.
- There are thousands of papers in journals on these topics

But very few of these books and papers actually provide formal definitions.

- When I try to read some of the papers, the definition seems to change from paragraph to paragraph
  - The notion is slippery!

Why should we care?

We're building autonomous agents that will need to make (moral) judgments

- Germany recently proposed a code for driverless cars. The proposal specified, among other things, that a driverless car should always opt for property damage over personal injury. Is this reasonable?

Why should we care?

We're building autonomous agents that will need to make (moral) judgments

- Germany recently proposed a code for driverless cars. The proposal specified, among other things, that a driverless car should always opt for property damage over personal injury. Is this reasonable?
  - Suppose that the probability of \$100,000 property damage is $.999$ and the probability of a minor injury is $.001$.
- A similar policy might preclude passing.
  - There's always a small risk of a personal injury ...

Can't we use Asimov's three laws?

Isaac Asimov, an American science fiction writer, proposed three laws of robotics in a short story written in 1942:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Can't we use Asimov's three laws?

Isaac Asimov, an American science fiction writer, proposed three laws of robotics in a short story written in 1942:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Unfortunately, it's not always clear how to apply these laws. Asimov's stories often considered conflicts that arose in trying to apply them.

- What if preventing harm to one person causes harm to another?
  - We need to think in terms of tradeoffs
- Should an assistive robot help an elderly patient commit suicide?

Asimov's Zeroth Law

An aside: Asimov later introduced a "zeroth law":

0. A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

This is relevant to concerns about super-intelligence ...

The Trolley Problem

The trolley problem was introduced by Philippa Foot [1967] and then examined carefully by Judith Thomson [1972] and many, many others:

> *Suppose that a runaway trolley is heading down the tracks. There are 5 people tied up on the track, who cannot move. If the trolley continues, it will kill all 5 of them. While you cannot stop the trolley, you can pull a lever, which will divert it to a side track. Unfortunately, there is a man on the side track who will get killed if you pull the lever. What is appropriate thing to do here? What is your degree of moral responsibility for the outcome if you do/do not pull the lever.*

The Trolley Problem

The trolley problem was introduced by Philippa Foot [1967] and then examined carefully by Judith Thomson [1972] and many, many others:

> *Suppose that a runaway trolley is heading down the tracks. There are 5 people tied up on the track, who cannot move. If the trolley continues, it will kill all 5 of them. While you cannot stop the trolley, you can pull a lever, which will divert it to a side track. Unfortunately, there is a man on the side track who will get killed if you pull the lever. What is appropriate thing to do here? What is your degree of moral responsibility for the outcome if you do/do not pull the lever.*

- Would you feel differently about throwing a fat man off the bridge to stop the train?

A modern version of the trolley problem [The social dilemma of autonomous vehicles, Bonnefon, Sharif, Rahwan, *Science* 2016]:

Should an autonomous vehicle swerve and kill its passenger when otherwise it would kill 5 pedestrians?

A modern version of the trolley problem [The social dilemma of autonomous vehicles, Bonnefon, Sharif, Rahwan, *Science* 2016]:

Should an autonomous vehicle swerve and kill its passenger when otherwise it would kill 5 pedestrians?

- People thought it should, but wouldn't buy an autonomous vehicle programmed this way!
- Note here that we have a conflict in Asimov's Laws of Robotics.

Moral Responsibility

There seems to be general agreement that moral responsibility involves *causality*,

- Agent a can't be morally responsible for outcome O if a's action didn't cause O.

Moral Responsibility

There seems to be general agreement that moral responsibility involves *causality*,

- Agent a can't be morally responsible for outcome O if a's action didn't cause O.

some notion of *blameworthiness*,

- To what extent is a to blame for outcome O?
- What could a have done to prevent O from happening?
  - What were a's alternatives?

Moral Responsibility

There seems to be general agreement that moral responsibility involves *causality*,

- Agent a can't be morally responsible for outcome O if a's action didn't cause O.

some notion of *blameworthiness*,

- To what extent is a to blame for outcome O?
- What could a have done to prevent O from happening?
  - What were a's alternatives?

and *intent*

- Did a want O to happen, or was O an unintended byproduct of a's real goal.
  - In the trolley problem, a didn't intend the person on the side track to die; he just wanted to save the 5 people on the main track

Moral Responsibility

There seems to be general agreement that moral responsibility involves *causality*,

- Agent a can't be morally responsible for outcome O if a's action didn't cause O.

some notion of *blameworthiness*,

- To what extent is a to blame for outcome O?
- What could a have done to prevent O from happening?
  - What were a's alternatives?

and *intent*

- Did a want O to happen, or was O an unintended byproduct of a's real goal.
  - In the trolley problem, a didn't intend the person on the side track to die; he just wanted to save the 5 people on the main track
- Not everyone agrees that intent is relevant
  - although people do seem to take it into account when judging moral responsibility

Causality

The literature considers two flavors of causality:

- *type causality*: smoking causes cancer
- *token/actual causality*: the fact that Willard smoked for 30 years caused him to get cancer

I have focused on token causality.

- The basic idea: counterfactuals:
  - A is a cause of B if, had A not happened, B wouldn't have happened.
  - *But-for causality*: the definition used in the law

It's not that easy:

> *[Lewis:] Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle if Suzy's throw had not preempted it.*

We want to call Suzy a cause of the bottle shattering, not Billy

- But even if Suzy hadn't thrown, the bottle would have shattered

There has been *lots* of work on getting good models of causality.

- Key influential recent idea: use *structural equations* to model the effect of interventions

Structural-equations models for causality

Idea: [Pearl] World described by variables that affect each other

- This effect is modeled by *structural equations*.

Split the random variables into

- *exogenous* variables
  - values are taken as given, determined by factors outside model
- *endogenous* variables.

Structural equations describe the values of endogenous variables in terms of exogenous variables and other endogenous variables.

- Have an equation for each variable
  - $X = Y + U$ does not mean $Y = U - X$!

Example 1: Arsonists

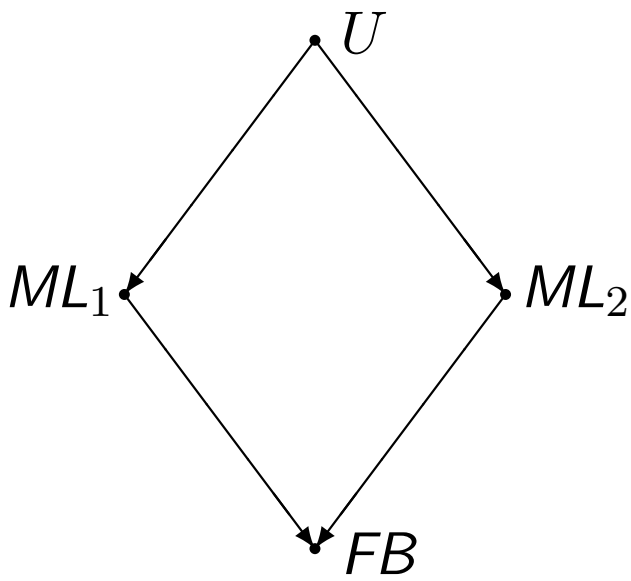
Two arsonists drop lit matches in different parts of a dry forest, and both cause trees to start burning. Consider two scenarios.

1. Disjunctive scenario: either match by itself suffices to burn down the whole forest.
2. Conjunctive scenario: both matches are necessary to burn down the forest

We can describe these scenarios using a *causal network*, whose nodes are labeled by the variables.

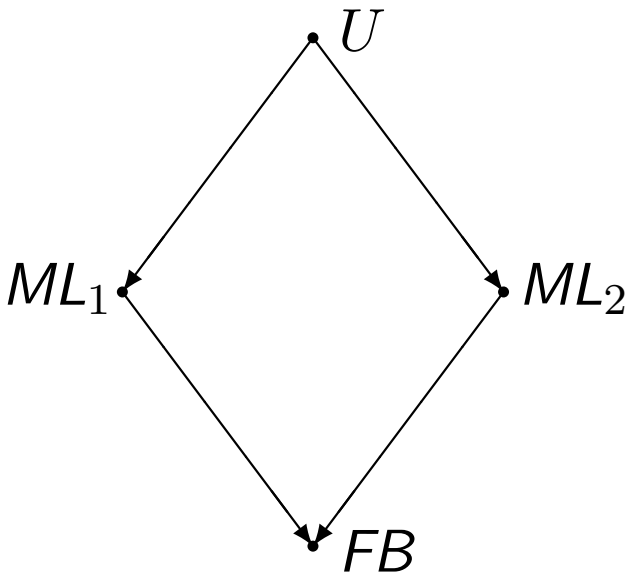
Arsonist Scenarios

Same causal network for both scenarios:



- endogenous variables ML_i, $i = 1, 2$:
  - $ML_i = 1$ iff arsonist i drops a match
- exogenous variable $U = (j_1 j_2)$
  - $j_i = 1$ iff arsonist i intends to start a fire.
- endogenous variable FB (forest burns down).
  - For the disjunctive scenario $FB = ML_1 \vee ML_2$
  - For the conjunctive scenario $FB = ML_1 \wedge ML_2$

Causal Networks

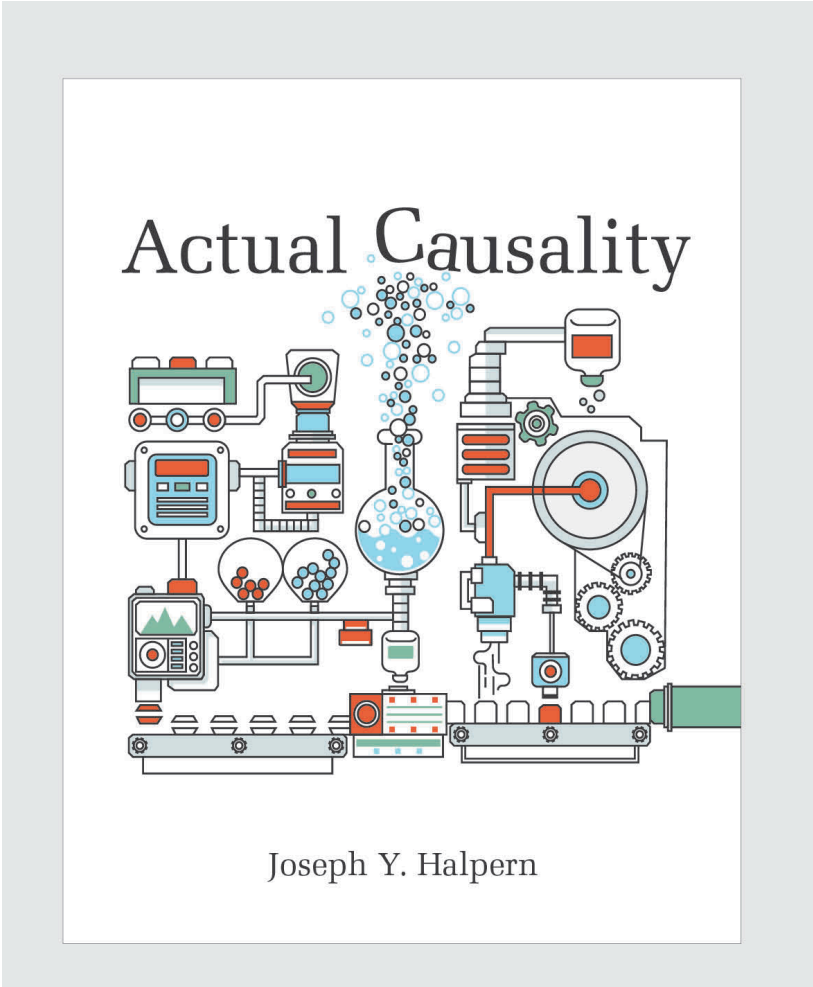


In a causal network, the arrows determine the "flow" of causality:

- There is an arrow from A to B if the equation for B depends on the value of A.
- The exogenous variables are at the top
- We restrict to scenarios where the causal network is *acyclic*: no cycles of influence
  - That means that, once we set the exogneous variables, we can determine the values of all the endogenous variables.

Judea Pearl and I gave a definition of causality using structural equations to model counterfactuals and the effect of interventions.

- If you're interested in the details . . .



Uncertainty

The definition of causality is relative to a *setting* (M, u)

- M is the causal model
  - Describes the variables and equations
- u is the *context* (i.e., what actually happened)
  - which arsonists dropped the match

Uncertainty

The definition of causality is relative to a *setting* (M, u)

- M is the causal model
  - Describes the variables and equations
- u is the *context* (i.e., what actually happened)
  - which arsonists dropped the match

In general, an agent has uncertainty about the true setting:

- Is one match enough to start the fire, or do we need two?
- Did the other arsonist drop the match

So we assume that the agent has a probability $\Pr$ on settings.

Uncertainty

The definition of causality is relative to a *setting* (M, u)

- M is the causal model
  - Describes the variables and equations
- u is the *context* (i.e., what actually happened)
  - which arsonists dropped the match

In general, an agent has uncertainty about the true setting:

- Is one match enough to start the fire, or do we need two?
- Did the other arsonist drop the match

So we assume that the agent has a probability $\Pr$ on settings.

Because of this uncertainty, an agent doesn't know whether performing an action ACT will actually cause an outcome O.

- ACT may cause O in some settings, but not in others.

But a can compute the probability that ACT causes O.

Degree of Blameworthiness

a can also compute the effect on outcome O of switching from action ACT to ACT':

- The switch may have no effect on O
- It may change the outcome away from O
- It may result in O in cases O wouldn't have happened

Let $\mathit{diff}(ACT, ACT')$ be the net change in the probability of outcome O happening if we switch from ACT to ACT':

$$\mathit{diff}(ACT, ACT', O) =$$
$$\Pr(ACT' \text{ does not result in } O \text{ but } ACT \text{ does})$$
$$-\Pr(ACT \text{ does not result in } O \text{ but } ACT' \text{ does})$$

Degree of Blameworthiness

a can also compute the effect on outcome O of switching from action ACT to ACT':

- The switch may have no effect on O
- It may change the outcome away from O
- It may result in O in cases O wouldn't have happened

Let $\mathit{diff}(ACT, ACT')$ be the net change in the probability of outcome O happening if we switch from ACT to ACT':

$$\begin{aligned}&\mathit{diff}(ACT, ACT', O) = \\ &\quad \Pr(ACT' \text{ does not result in } O \text{ but } ACT \text{ does}) \\ &\quad -\Pr(ACT \text{ does not result in } O \text{ but } ACT' \text{ does})\end{aligned}$$

The *degree of blameworthiness* of ACT for O is the largest net change (over all actions that a can perform):

$$db(ACT, O) = \max_{ACT'} \mathit{diff}(ACT, ACT', O)$$

- Intuitively, $db(ACT, O)$ measures the extent to which performing an action other than ACT can affect outcome O.

Some Subtleties

The degree of blameworthiness depends on the probability $\Pr$ on settings

Example: To what extent is one of the arsonists to blame for the forest fire?

- It depends on
  - how likely is the conjunctive vs. disjunctive scenario?
  - how likely the other arsonist is to drop the match?

Some Subtleties

The degree of blameworthiness depends on the probability $\Pr$ on settings

Example: To what extent is one of the arsonists to blame for the forest fire?

- It depends on
  - how likely is the conjunctive vs. disjunctive scenario?
  - how likely the other arsonist is to drop the match?

Suppose each arsonist thinks that (with high probability) we are in the disjunctive scenario and that the other arsonist will drop a match.

- Then each has low degree of blameworthiness.
- Nothing either one could do would have made a difference

Some Subtleties

The degree of blameworthiness depends on the probability $\Pr$ on settings

Example: To what extent is one of the arsonists to blame for the forest fire?

- It depends on
  - how likely is the conjunctive vs. disjunctive scenario?
  - how likely the other arsonist is to drop the match?

Suppose each arsonist thinks that (with high probability) we are in the disjunctive scenario and that the other arsonist will drop a match.

- Then each has low degree of blameworthiness.
- Nothing either one could do would have made a difference
- But between them they caused the fire!

Some Subtleties

The degree of blameworthiness depends on the probability $\Pr$ on settings

Example: To what extent is one of the arsonists to blame for the forest fire?

- It depends on
  - how likely is the conjunctive vs. disjunctive scenario?
  - how likely the other arsonist is to drop the match?

Suppose each arsonist thinks that (with high probability) we are in the disjunctive scenario and that the other arsonist will drop a match.

- Then each has low degree of blameworthiness.
- Nothing either one could do would have made a difference
- But between them they caused the fire!

Although each individual has low degree of blameworthiness, the group has degree of blameworthiness 1.

- This is like the tragedy of the commons.
- No individual has much blame, but the group does

More subtleties

Example: Suppose that a doctor's use of a drug to treat a patient is the cause of a patient's death. But the doctor had no idea there would be adverse side effects. Then, according to his probability distribution (which we think of as representing his prior beliefs, before he treated the patient), his degree of blameworthiness is low.

More subtleties

Example: Suppose that a doctor's use of a drug to treat a patient is the cause of a patient's death. But the doctor had no idea there would be adverse side effects. Then, according to his probability distribution (which we think of as representing his prior beliefs, before he treated the patient), his degree of blameworthiness is low.

- But are the doctor's prior beliefs the right beliefs to use?
- What if there were articles in leading medical journals about the adverse effects of the drug?
- We can instead use the probability distribution that a reasonable conscientious doctor would have had.
  - The definition of blameworthiness is relative to a probability distribution.
    - The modeler needs to decide which probability distribution to use.

Tradeoffs

Blameworthiness is relative to an outcome.

- Depending on how he pulls the lever, the agent has degree of blameworthiness 1 for either the death of five people or the death of one person

Tradeoffs

Blameworthiness is relative to an outcome.

- Depending on how he pulls the lever, the agent has degree of blameworthiness 1 for either the death of five people or the death of one person

So what should he do?

- We can evaluate tradeoffs using a *utility function*
  - how much the agent values each outcomes
- Given a utility function, take the action ACT that maximizes the agent's *expected utility*
  - for each outcome O, multiply the probability of O occurring if ACT is performed by the utility of O

Tradeoffs

Blameworthiness is relative to an outcome.

- Depending on how he pulls the lever, the agent has degree of blameworthiness 1 for either the death of five people or the death of one person

So what should he do?

- We can evaluate tradeoffs using a *utility function*
  - how much the agent values each outcomes
- Given a utility function, take the action ACT that maximizes the agent's *expected utility*
  - for each outcome O, multiply the probability of O occurring if ACT is performed by the utility of O
- But which utility function should be used?
  - There is no "right" utility function, but we tend to view some as more reasonable than others.
    - It's OK to kill 5 people if the one you're saving is your child

Intention

Intuition: Agent a who performed ACT intended O if, had a been unable to impact O, a would not have performed ACT.

- In the trolley problem, the death of the person on the sidetrack was not intended; you would have pulled the lever in any case whether or not the man died)

We can make this precise using causal models and the agent's utility function.

Intention

Intuition: Agent a who performed ACT intended O if, had a been unable to impact O, a would not have performed ACT.

- In the trolley problem, the death of the person on the sidetrack was not intended; you would have pulled the lever in any case whether or not the man died)

We can make this precise using causal models and the agent's utility function.

We also need to deal with situations where an agent intends multiple outcomes.

- Example: An assassin plants a bomb to intending to kill two people. He would have planted it anyway if only one had died.

A definition that deals with all this is given in the paper.

Putting It All Together

Psychologists have done experiments to determine when an act is viewed as *morally acceptable*. A first cut:

- An action is morally acceptable if it maximizes the agent's expected utility, and the agent had "reasonable" probability and utility functions.

Putting It All Together

Psychologists have done experiments to determine when an act is viewed as *morally acceptable*. A first cut:

- An action is morally acceptable if it maximizes the agent's expected utility, and the agent had "reasonable" probability and utility functions.
  - The notion of "reasonable" can take into account the agent's computational limitations and his "emotional state" (age, recent events, . . .)

Putting It All Together

Psychologists have done experiments to determine when an act is viewed as *morally acceptable*. A first cut:

- An action is morally acceptable if it maximizes the agent's expected utility, and the agent had "reasonable" probability and utility functions.
  - The notion of "reasonable" can take into account the agent's computational limitations and his "emotional state" (age, recent events, . . .)
  - The agent can still be held blameworthy for some outcomes of his action, even if the action is morally acceptable, on this view.
- This clearly isn't enough to capture people's views.
  - People take intention into accunt.
  - People also compare seem to compare actions performed to default actions.
    - It's complicated!

Key points for a computer scientist:

- Given a probability and utility, degree of blameworthiness and intention can be computed efficiently.

Key points for a computer scientist:

- Given a probability and utility, degree of blameworthiness and intention can be computed efficiently.
- The probabilities can be determined from data.
- Can we give an autonomous agent "reasonable" utilities?
  - This is the "value alignment" problem
  - Just watching humans may not reveal moral behavior

Key points for a computer scientist:

- Given a probability and utility, degree of blameworthiness and intention can be computed efficiently.
- The probabilities can be determined from data.
- Can we give an autonomous agent "reasonable" utilities?
  - This is the "value alignment" problem
  - Just watching humans may not reveal moral behavior

These definitions don't solve the problem, but at least they can help make it clear what we're disagreeing about!

Final Words

- I have presumed that agents have a probability on settings
  - It's not clear that probability is always the best/most reasonable way to represent uncertainty.
    - Need to think about how to modify the definitions to deal with this other representations of uncertainty.
- I hasn't said anything about how we decide what counts as a reasonable/acceptable utility function.
  - It's doubtful that we can get universal agreement on that.
  - But we can and should try to reach some consensus, at least when it comes to the autonomous agents we implement

Final Words

- I have presumed that agents have a probability on settings
  - It's not clear that probability is always the best/most reasonable way to represent uncertainty.
    - Need to think about how to modify the definitions to deal with this other representations of uncertainty.
- I hasn't said anything about how we decide what counts as a reasonable/acceptable utility function.
  - It's doubtful that we can get universal agreement on that.
  - But we can and should try to reach some consensus, at least when it comes to the autonomous agents we implement
    - This is a task we all need to be involved in.

Issues of moral responsibility are subtle.

- As AI systems become more powerful, we'll need to confront them
  - Self-driving cars
  - AI systems making decisions on who gets loans, who goes on parole
  - robots in nursing homes
- We need an informed citizenry to make good decisions.
  - Don't leave it to the experts!