Blameworthiness and Intention: Towards Formal Definitions

Joe Halpern
Cornell University

Includes joint work with Meir Friedenberg (Cornell) (on group blameworthiness), Max Kleiman-Weiner (MIT/Harvard) (on single-agent blameworthiness and intention), and Judea Pearl (UCLA) (on causality).

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 ...

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!

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

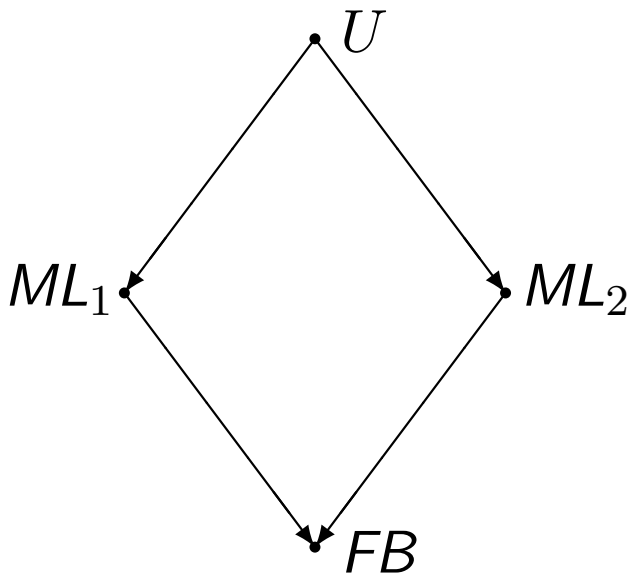
Two arsonists drop lit matches in different parts of a dry forest, and both cause trees to start burning. Consider two scenarios.

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.

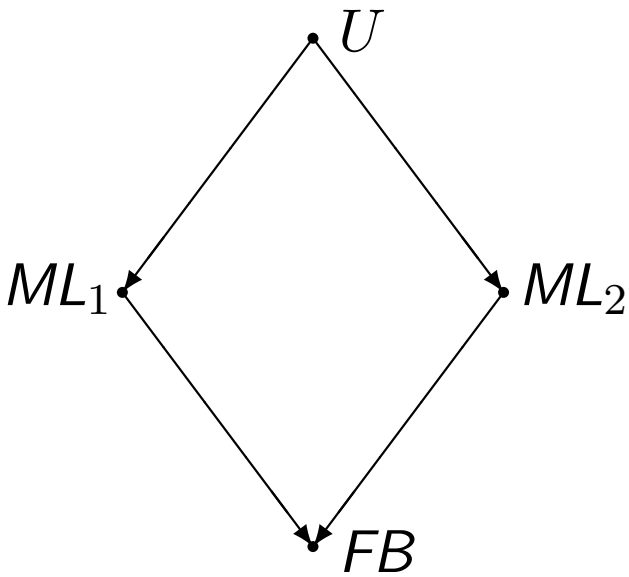
Arsonist scenarios

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

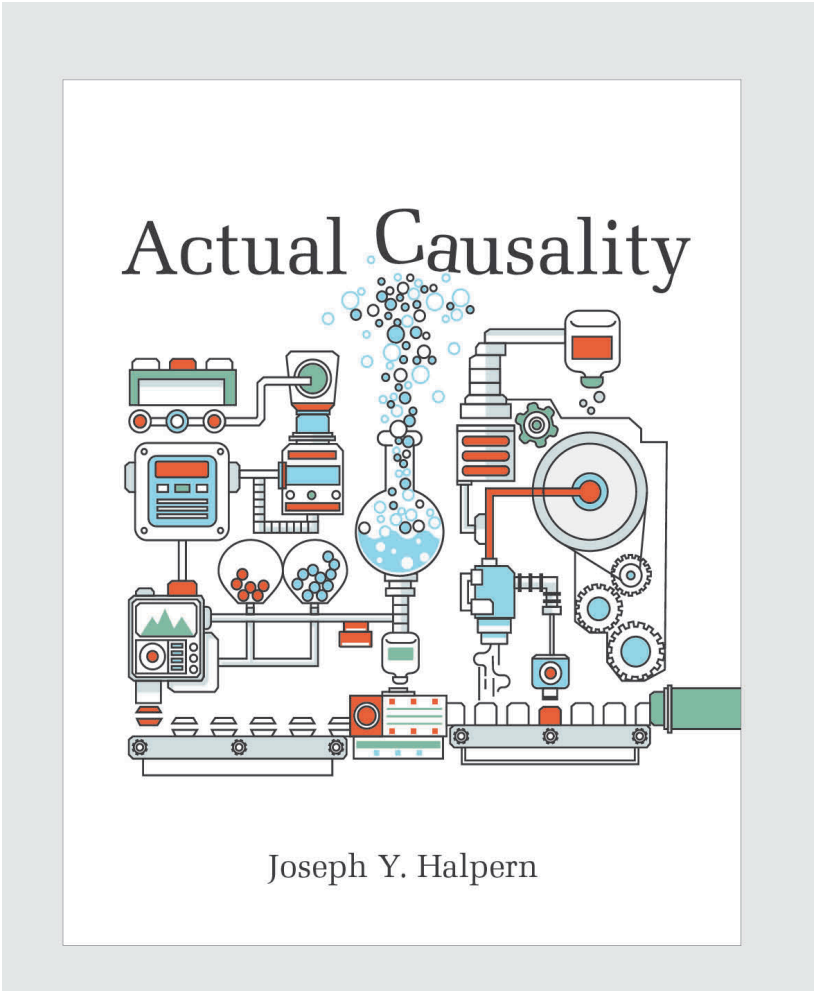


In a causal network, the arrows determine the "flow" of causality:

- 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.

- I won't say more about causality here. 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.

Net effect of an action

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

Net effect of an action

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 $\mathit{dff}(act, act')$ be the net change in the probability of outcome o happening if we switch from act to act':

$$\begin{aligned}&\mathit{dff}(act, act', o) = \\ &\quad \Pr(act' \text{ does not result in } o \text{ but } act \text{ does}) \\ &\quad -\Pr(act \text{ does not result in } o \text{ but } act' \text{ does})\end{aligned}$$

Net effect of an action

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 $\mathit{dff}(act, act')$ be the net change in the probability of outcome o happening if we switch from act to act':

$$\begin{aligned}&\mathit{dff}(act, act', o) = \\ &\quad \Pr(act' \text{ does not result in } o \text{ but } act \text{ does}) \\ &\quad -\Pr(act \text{ does not result in } o \text{ but } act' \text{ does})\end{aligned}$$

The *net effect* of act on o considers the largest net change (over all actions that a can perform):

$$\mathit{dff}(act, o) = \max_{act'} \mathit{dff}(act, act', o)$$

Net effect of an action

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 $\mathit{dff}(act, act')$ be the net change in the probability of outcome o happening if we switch from act to act':

$$\begin{aligned}&\mathit{dff}(act, act', o) = \\ &\quad \Pr(act' \text{ does not result in } o \text{ but } act \text{ does}) \\ &\quad -\Pr(act \text{ does not result in } o \text{ but } act' \text{ does})\end{aligned}$$

The *net effect* of act on o considers the largest net change (over all actions that a can perform):

$$\mathit{dff}(act, o) = \max_{act'} \mathit{dff}(act, act', o)$$

- Intuitively, $\mathit{dff}(act, o)$ measures the extent to which performing an action other than act can affect outcome o.

Some Subtleties I: Cost

There seems to be more to to blameworthiness than net effect:
Example: Suppose that Bob could have given up his life to save Tom. Bob decided to do nothing, so Tom died. The difference between the probability of Tom dying if Bob does nothing and if Bob gives up his life is 1.

- The net effect of Bob's doing nothing on Tom's death is 1.

Yet people are not so inclined to blame Bob.

- Intuitively, the *cost* of saving Tom is too high

Degree of blameworthiness definition

We capture this by associating with every action a its cost, $c(a)$.

- In defining degree of blameworthiness, we combine the cost with the net effect of a.
  - How much we weight the cost is captured by a parameter N

Degree of blameworthiness definition

We capture this by associating with every action a its cost, $c(a)$.

- In defining degree of blameworthiness, we combine the cost with the net effect of a.
  - How much we weight the cost is captured by a parameter N
- $\lim_{N\to\infty} \mathit{db}_M(act, act', \varphi) = \mathit{diff}(act, act', \varphi)$.
  - As N gets large, we essentially ignore the costliness of the act.
- if N and $c(act')$ are both close to $\max_{act''} c(act'')$ and $c(act) = 0$, then $\mathit{db}_N(act, act', \varphi)$ is close to 0.
  - In the example with Bob and Tom, if we take costs seriously, then we do not find Bob particularly blameworthy.

In general, we expect the choice of N to be situation-dependent.

Degree of blameworthiness definition

We capture this by associating with every action a its cost, $c(a)$.

- In defining degree of blameworthiness, we combine the cost with the net effect of a.
  - How much we weight the cost is captured by a parameter N
- $\lim_{N\to\infty} \mathit{db}_M(act, act', \varphi) = \mathit{diff}(act, act', \varphi)$.
  - As N gets large, we essentially ignore the costliness of the act.
- if N and $c(act')$ are both close to $\max_{act''} c(act'')$ and $c(act) = 0$, then $\mathit{db}_N(act, act', \varphi)$ is close to 0.
  - In the example with Bob and Tom, if we take costs seriously, then we do not find Bob particularly blameworthy.

In general, we expect the choice of N to be situation-dependent.

The *degree of blameworthiness of* act *for* φ *relative to* act' *(given* c *and* N*)* is

$$\mathit{db}_N(act, act', \varphi) = \mathit{diff}(act, act', \varphi)\frac{N - \max(c(act') - c(act), 0)}{N}.$$

The degree of blameworthiness of act for φ is

$$\mathit{db}_N(act, \varphi) = \max_{act'} \mathit{db}_N(act, act', \varphi).$$

Some Subtleties II: Group Blameworthiness

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 II: Group Blameworthiness

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 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 II: Group Blameworthiness

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 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 II: Group Blameworthiness

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 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 plausibly has degree of blameworthiness 1.

- This is like the tragedy of the commons.
- I return to group blameworthiness later in the talk

Some subtleties III: Whose probability?

Example: Suppose that a drug used by a doctor to treat a patient caused the patient's death. 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.

Some subtleties III: Whose probability?

Example: Suppose that a drug used by a doctor to treat a patient caused the patient's death. 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 must decide which probability distribution to use.

Tradeoffs

Blameworthiness is relative to an outcome.

- Depending on how he pulls the lever, the agent has some degree of blameworthiness 1 for either the death of five people or the death of one person
  - The exact degree of blameworthiness depends on the cost of each alternative.
  - People might well describe a lower blameworthiness to someone who kills 1 rather than 5, and even less blameworthiness to someone who refuses to push a fat man off a bridge

But there is more to judgments of moral responsibility than blameworthiness . . .

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.

Back to group blameworthiness

Clearly in the disjunctive arsonist scenario, the two arsonists together were responsible for the forest burning down.

- It seems that each agent should bear some blame for the outcome as a member of the group.
  - But what if the arsonists didn't know each other? Is that different from the case where they planned together to drop the matches?
- We believe that the ability of the agents to coordinate must be taken into account when ascribing blame.
- Well instead define blameworthiness in multi-agent settings by:
  1. Defining a notion of the blameworthiness of a group (that takes coordination into account), and
  2. Showing how to distribute group blameworthiness among member agents.

Ascribing group blameworthiness

The definition of group blameworthiness is similar in spirit to the earlier definition of blameworthiness for a single agent.

- we compute an analogue of $diff(act, act', \varphi)$, but instead of comparing two acts, we compare two *probabilities* $\Pr$ and $\Pr'$
  - Intuitively, $\Pr$ is the initial distribution over causal settings; $\Pr'$ is a new distribution over causal settings that some coordination actions among the agents can bring about
  - This lets us ignore the details of a much richer (and likely quite complicated) causal model that captures possible negotiations, discussions, and actions by the agents
- The cost function now takes into account of bringing about distribution $\Pr'$

Apportioning group blameworthiness among agents

Suppose we can ascribe blame to each subgroup of agents.

- We want a way of apportioning the group blame to members of the group

Apportioning group blameworthiness among agents

Suppose we can ascribe blame to each subgroup of agents.

- We want a way of apportioning the group blame to members of the group

We take an axiomatic approach, and consider three axioms that the ascription of individual blame should satisfy:

- *Efficiency:* The sum of the blames of each individual agent should be the blame ascribed to the group of all agents.
  - This capture the intuition that we are trying to apportion the blame.
- *Symmetry*: The names of agents should not affect their blameworthiness, so if we simply rename them then the blameworthiness ascribed to them should remain the same.
- *Strong Monotonicity:* If agent j contributes more to the group blameworthiness of all groups in one scenario than another, then j also ought to have a greater degree of (personal) blameworthiness in the first scenario.

Apportioning group blameworthiness among agents

Suppose we can ascribe blame to each subgroup of agents.

- We want a way of apportioning the group blame to members of the group

We take an axiomatic approach, and consider three axioms that the ascription of individual blame should satisfy:

- *Efficiency*
- *Symmetry*
- *Strong Monotonicity*

Young [1985] showed that the only distribution procedure that would satisfy Efficiency, Symmetry, and Strong Monotonicity is the *Shapley value*.

Ferey and Dehez (2016) independently applied the Shapley value in sequential-liability cases, with the same intuition (though to apportioning restitution of damages, rather than blameworthiness).

- They also showed that this approach can explain examples from the case law and aligns with a systematization of liability ascription laid out in the legal literature.

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.
  - There are many theories of moral acceptability that reject maximizing expected utility
  - People take intention into account.
  - People also 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

- It would be useful to have psychology experiments to determine to what extent these definitions are compatible with how people ascribe blameworthines
  - E.g., Do they take cost into account? If so, how?

Final Words

- It would be useful to have psychology experiments to determine to what extent these definitions are compatible with how people ascribe blameworthines
  - E.g., Do they take cost into account? If so, how?
- 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 other representations of uncertainty.
    - Would different representations lead to qualitatively different results?

Final Words

- It would be useful to have psychology experiments to determine to what extent these definitions are compatible with how people ascribe blameworthines
  - E.g., Do they take cost into account? If so, how?
- 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 other representations of uncertainty.
    - Would different representations lead to qualitatively different results?
- I haven'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

- It would be useful to have psychology experiments to determine to what extent these definitions are compatible with how people ascribe blameworthines
  - E.g., Do they take cost into account? If so, how?
- 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 other representations of uncertainty.
    - Would different representations lead to qualitatively different results?
- I haven'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.

References

- M. Friedenberg and J. Y. Halpern, Blameworthiness in multi-agent settings, AAAI '19
- J. Y. Halpern and M. Kleiman-Weiner, Towards formal definitions of blameworthiness, intention, and moral responsibility, AAAI '18.