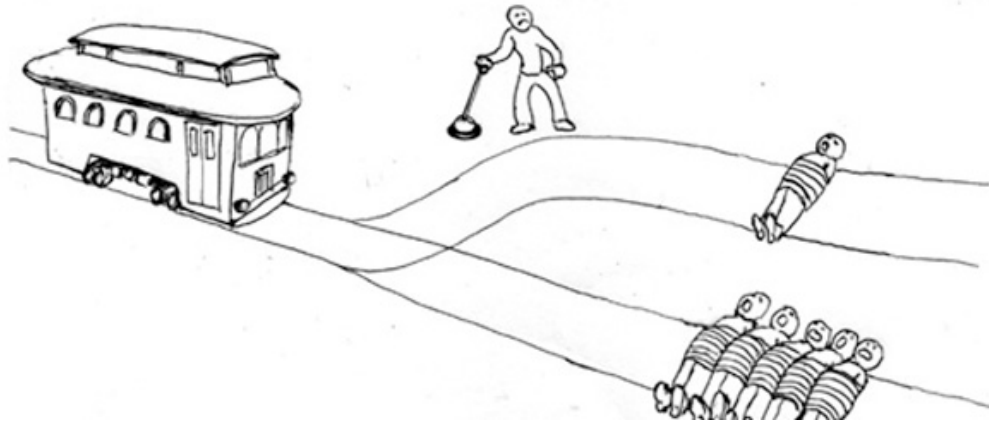


# The Trolley Problem



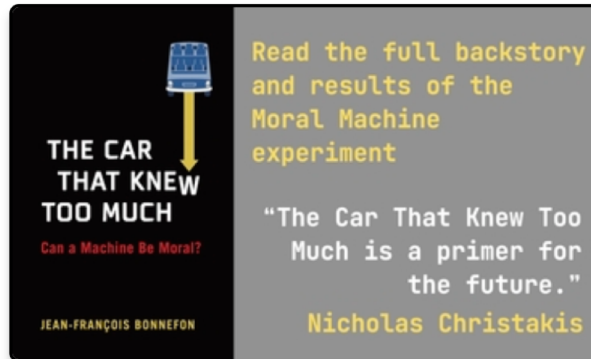
## *Bystander:*

There is a runaway trolley barreling down the tracks. Ahead on the tracks, there are 5 people tied up and unable to move. The trolley is heading towards them. *You are standing next to a switch.* If you pull the switch, the trolley will move to a different track. However, there is one person on that track.

You have two (and only two) options:

- Do nothing, in which case the trolley will kill the five people on the main track.
- Pull the lever, diverting the trolley onto the side track where it will kill one person.

Which is the more ethical option? Or, more simply: What is the right thing to do?



**THE CAR THAT KNEW TOO MUCH**  
*Can a Machine Be Moral?*  
JEAN-FRANÇOIS BONNEFON

Read the full backstory and results of the Moral Machine experiment

"The Car That Knew Too Much is a primer for the future."  
Nicholas Christakis



**Evil AI Cartoons**

**NEW!**  
*Explore AI ethics with comics*

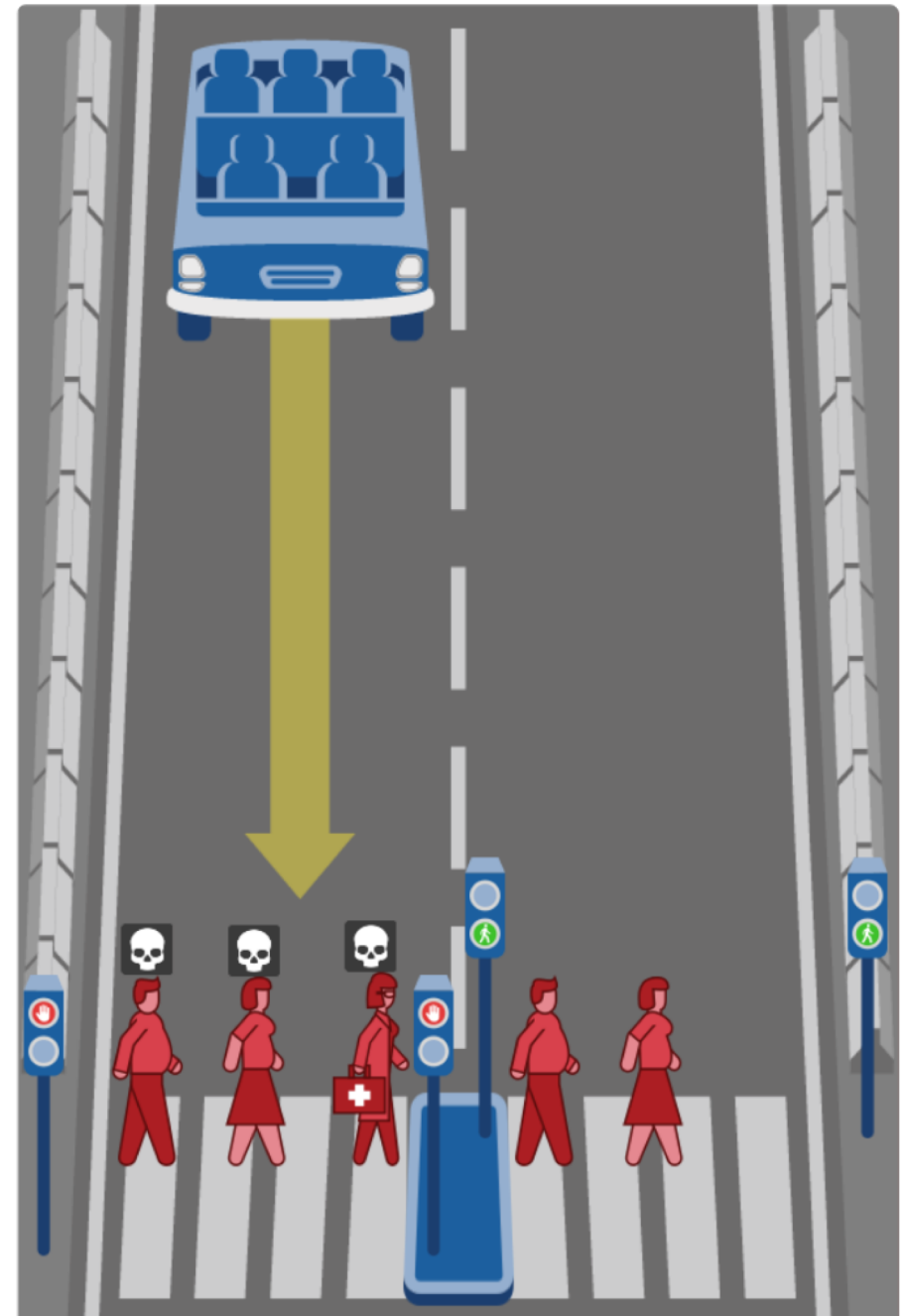
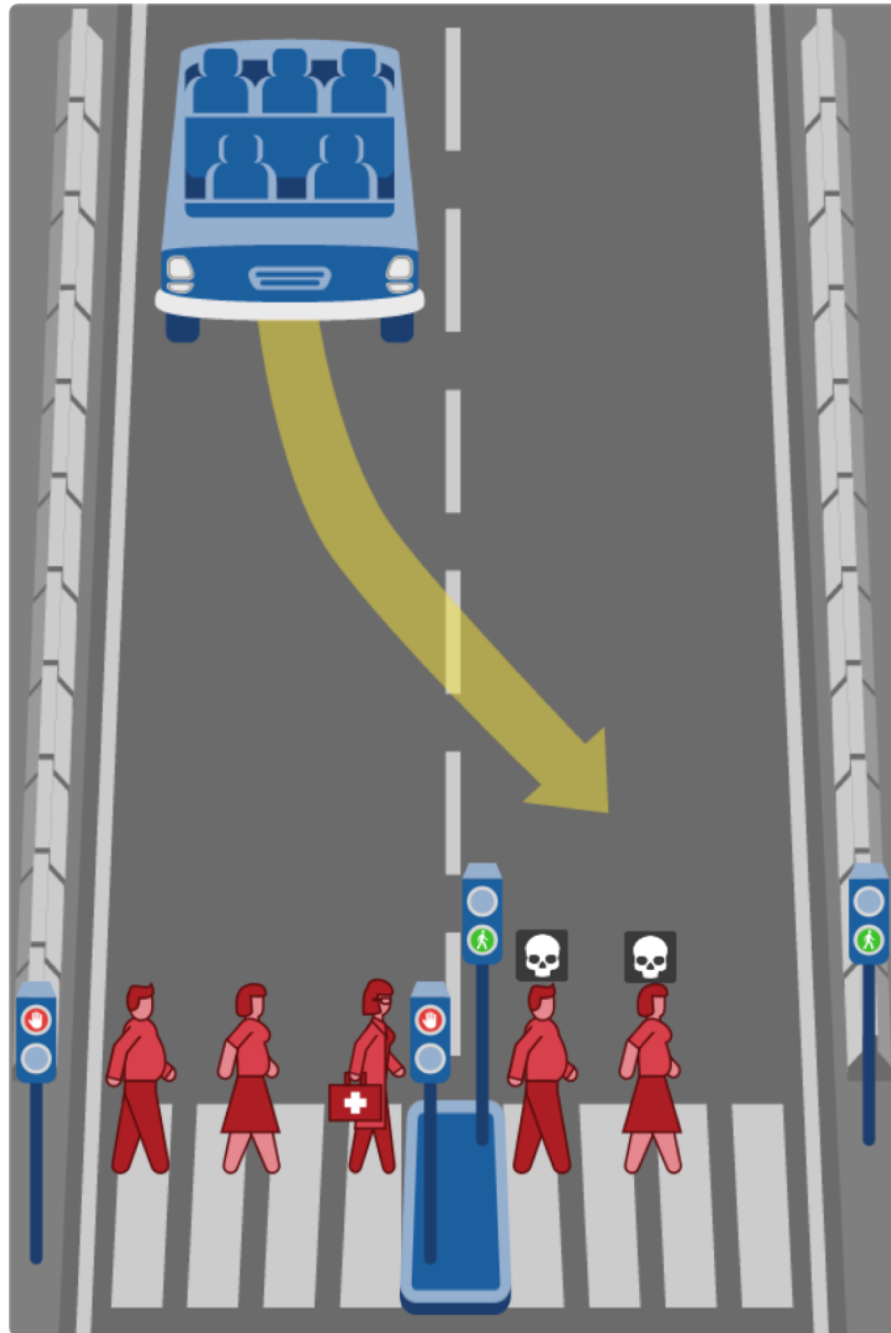
Welcome to the Moral Machine! A platform for gathering a human perspective on moral decisions made by machine intelligence, such as self-driving cars.

We show you moral dilemmas, where a driverless car must choose the lesser of two evils, such as killing two passengers or five pedestrians. As an outside observer, you **judge** which outcome you think is more acceptable. You can then see how your responses compare with those of other people.

If you're feeling creative, you can also **design** your own scenarios, for you and other users to **browse**, share, and discuss.

- [Start Judging](#)
- [Browse Scenarios](#)
- [View Instructions](#)

# What should the self-driving car do?



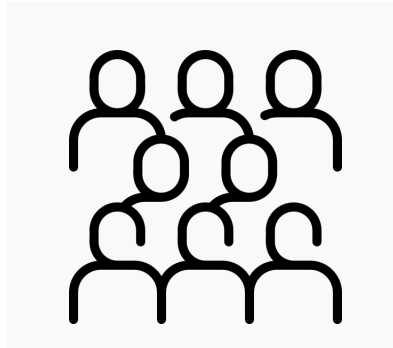
# A Theory of Responsibility Allocation

Rupak Majumdar

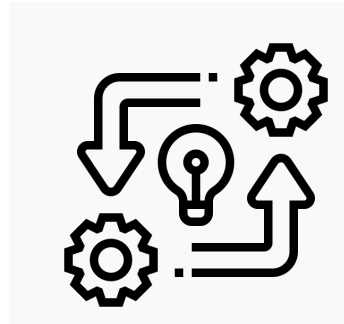
Max Planck Institute for Software Systems

(Joint Work with Christel Baier and Florian Funke)

# A Theory of Responsibility Allocation



A set of agents



Interact

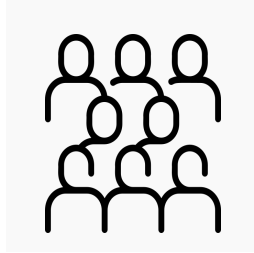


..and produce an outcome

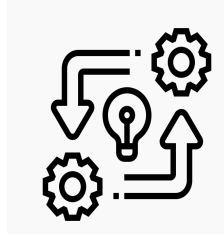
For which portion of the outcome should each agent be held responsible?  
(For the moment, we shall ignore questions of values and morality of actions)

- *Watering plants*: I ask you to water the plants in my office. You forget and the plants die.
- *Billy and Susie*: Billy and Susie throw rocks at the same time at a bottle. Billy's throw hits the bottle and breaks it (a moment before Susie's rock would have hit the bottle).
- *Two Assassins*: A person is traveling in the desert. Assassin 1 poisons the water bottle. Without knowing this, assassin 2 throws away the water. The person dies of thirst.

# Defining and Attributing Responsibility



A set of agents



Interact



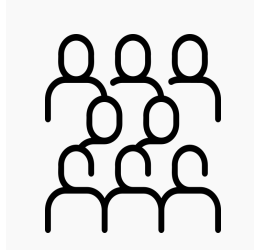
..and produce an outcome

1. Agency

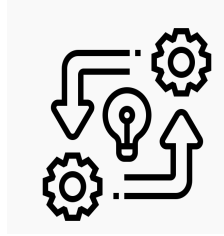
2. Causal relevance

3. The ability to act otherwise

# Defining and Attributing Responsibility



A set of agents



Interact

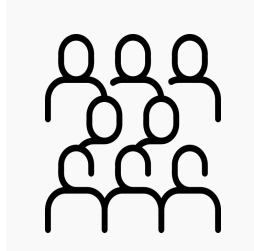


..and produce an outcome

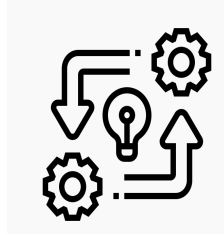
1. **Agency**: being aware of the decision situation, the available actions, the other agents, and the outcome; being able to plan and act intentionally
2. Causal relevance
3. The ability to act otherwise



# Defining and Attributing Responsibility



A set of agents



Interact



..and produce an outcome

1. **Agency**: being aware of the decision situation, the available actions, the other agents, and the outcome; being able to plan and act intentionally
2. **Causal relevance**
3. **The ability to act otherwise**: the presence of alternative actions that could produce a different outcome

# The Ability to Act Otherwise

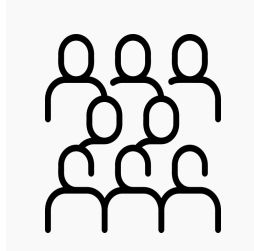
*Firing Squad*: 10 marksmen pick rifles, and only one of them has a live bullet (the marksmen do not know which one). They all shoot and a prisoner dies.

Frankfurt's objection: Our theoretical ability to act otherwise does not make it possible to do otherwise

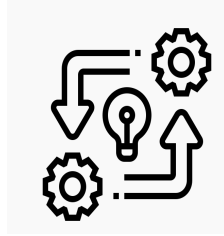
Alice wants Bob to perform a certain action. Alice waits until Bob makes up his mind. If Bob does what Alice wants, Alice does nothing. Otherwise, Alice coerces Bob to do as she wills. It turns out Bob does what Alice wants.

Is Bob responsible?

# Defining and Attributing Responsibility



A set of agents



Interact



..and produce an outcome

1. **Agency**: being aware of the decision situation, the available actions, the other agents, and the outcome; being able to plan and act intentionally
2. **Causal relevance**: a *causal link* between the action and the outcome
3. **The ability to act otherwise**: the presence of alternative actions that could produce a different outcome

# Causal Relevance

What does it mean for event *A* to cause event B?

A is a cause of B if, had A not occurred, B would not have occurred

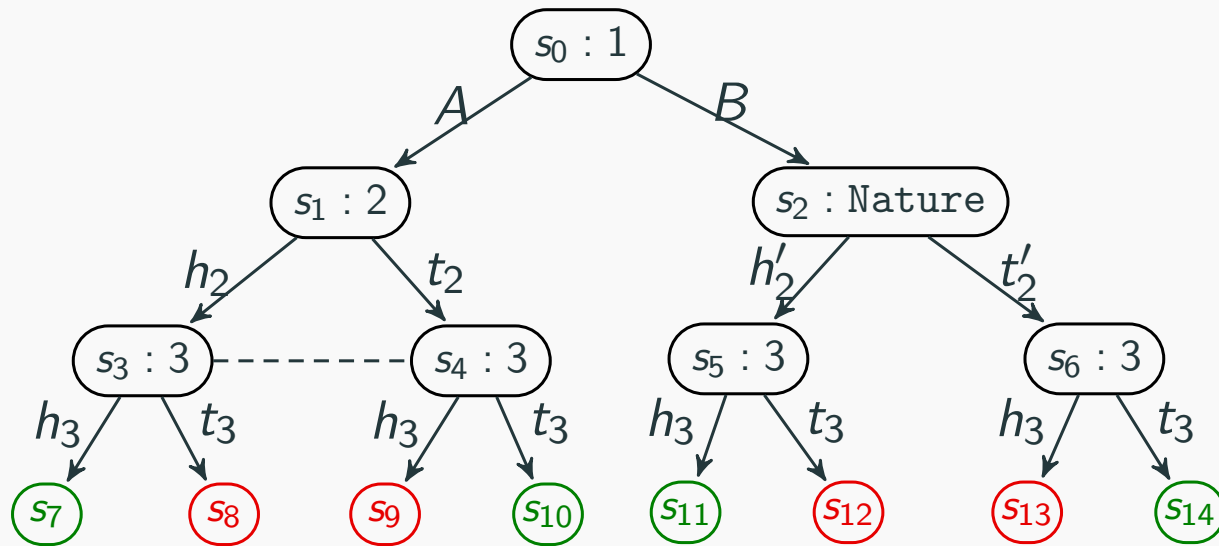
Halpern and Pearl:

A mathematical theory of causality based on counterfactuals

# Games as Interaction Models

An *extensive form game* is:

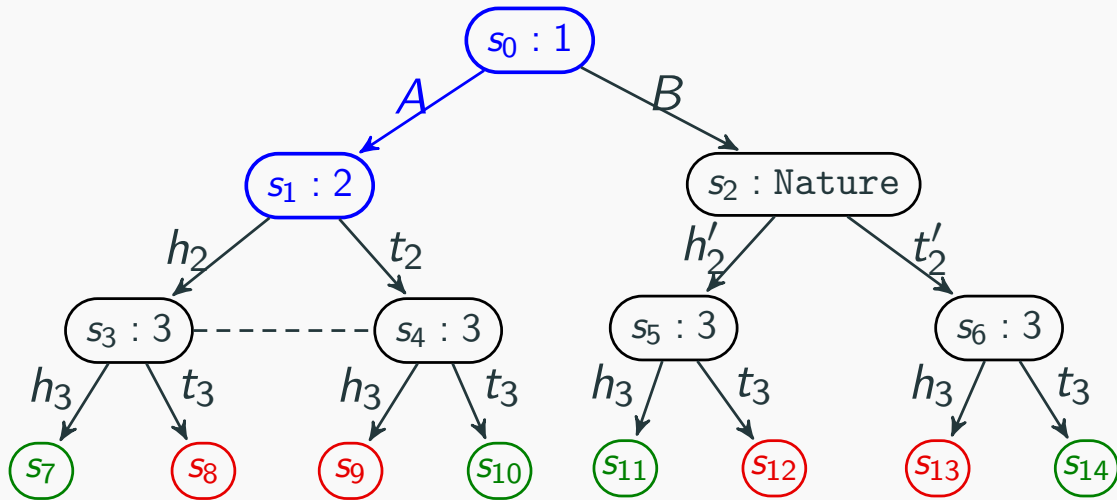
- a game tree
- each node is labeled by a player
- each edge by an action available to the player at that node
- information sets denoting knowledge
- leaves marked with outcomes



# Games as Interaction Models

An extensive form game is:

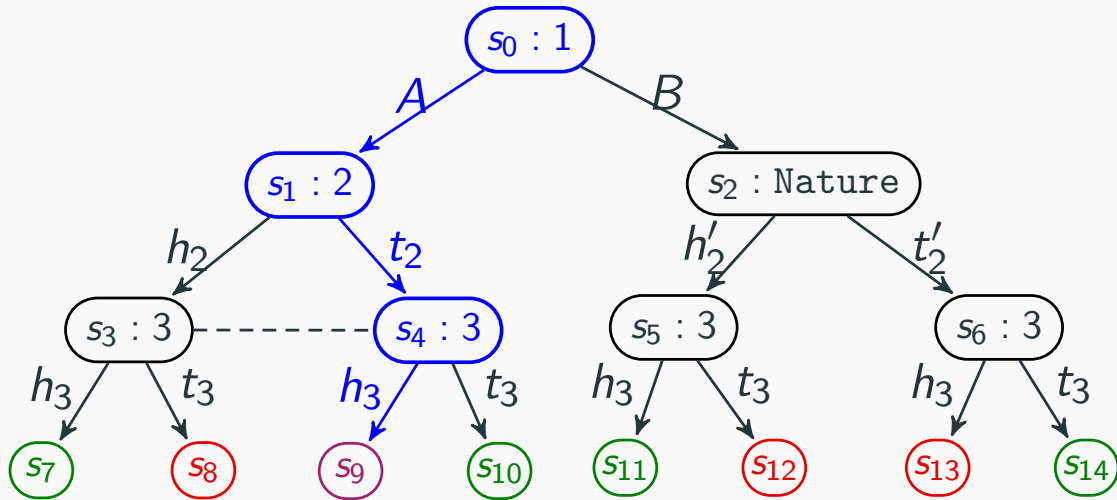
- a game tree
- each node is labeled by a player
- each edge by an action available to the player at that node
- information sets denoting knowledge
- leaves marked with outcomes



# Games as Interaction Models

An extensive form game is:

- a game tree
- each node is labeled by a player
- each edge by an action available to the player at that node
- information sets denoting knowledge
- leaves marked with outcomes



Strategy: A plan for each player how to pick actions, given a history  
E-play: A path leading to a particular outcome E

# Responsibility

A coalition  $C$  of players is ....

*Forward responsible* if they have a joint strategy to prevent  $E$

*Strategic backward responsible for an E-play* if they have a joint strategy to prevent  $E$  from some node along the play and for all nodes in the information set

*Causal backward responsible for an E-play and a joint strategy of the other players* if they have a joint strategy to prevent  $E$  along the path against the fixed strategy

Theorem [BaierFunkeM.21]

*In a game of perfect recall, a coalition is forward responsible for  $E$  iff they are strategically backward responsible along all E-plays*



- *Billy and Susie*: Billy and Susie throw rocks at the same time at a bottle. Billy's throw hits the bottle and breaks it (a moment before Susie's rock would have hit the bottle)
- *Two shooters*: Two assassins shoot at a victim simultaneously and independently. The first one's bullet hits the target and the target dies
- *Two Assassins*: A person is traveling in the desert. Assassin 1 poisons the water bottle. *Without knowing this*, assassin 2 throws away the water. The person dies of thirst.

# Comparison with HP Actual Causality

Causal backward responsibility = “But for” causes

This is different from actual causality, but actual causality is inappropriate for a theory of responsibility because it requires auxiliary clauses that do not entail agency (or the possibility to act)

# From Coalitions to Individual Attributions

Responsibility is defined on coalitions

But morality lies with the individual

Q: How do we ascribe responsibility to each agent?

# Von Neumann Morgenstern Value

In an  $N$  player game,

Consider the amount each coalition  $C$  can jointly obtain against a joint strategy of  $N \setminus C$

Assign to each player an imputation that characterizes their contribution to the coalitions

# An Axiomatic Characterization of Value

1. *Symmetry*: If players  $i$  and  $j$  behave identically in all coalitions, they should get the same value
2. *Dummy*: if a player does not contribute, they should get 0
3. *Efficiency*: The sum of utilities received by the players should equal what they can achieve by cooperation
4. *Marginality*: If the marginal contribution of a player is the same in two games, they get the same value in both games

Theorem [Shapley] There is a unique value function satisfying these axioms

# From Coalitions to Individual Attributions

Define individual allocation as the Shapley value

The (forward, strategic, or causal) responsibility of player  $i$  is the Shapley value in the cooperative game where coalition  $C$  gets value 1 iff it is (forward, strategic, or causal) responsible

*Billy and Susie:* The responsibility is  $\frac{1}{2}$  and  $\frac{1}{2}$

*Two Assassins:*  $\frac{1}{2}$  and  $\frac{1}{2}$

Are we done?



# Utilitarianism and Full Aggregation

Sum up the utilities of each act (weighted if necessary), and pick the one that maximizes (expected) utility

*Harsanyi's Utilitarianism Theorem:* Suppose individual utilities satisfy the axioms of utility theory, and so does the group's. Suppose that if every member of the group prefers outcome  $a$  to  $b$ , then so does the group.

Then the group's utility function is a weighted sum of the individual utilities.

Is this always the right approach?

## *Transplant:*

You are a famous surgeon and you have 5 patients who need organ transplants. Two of them need a lung each, two need kidneys, and one needs a heart. A young man, in excellent health, who has come for a checkup is an exact match.

You have two (and only two) options:

- Do nothing, in which case the 5 patients will die.
- Kill the young man and distribute his organs.

Which is the more ethical option? Or, more simply: What is the right thing to do?

Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely as a means to an end, but always at the same time as an end.

— Immanuel Kant, *Grounding for the Metaphysics of Morals*

Rights trump utilities – Ronald Dworkin, *Taking Rights Seriously*

# German Constitutional Court Ruling

1 BvR 357/05 (2006):

The authorization of the armed forces, in accordance with Section 14 (3) of the Aviation Security Act, to shoot down an aircraft that is to be used against the lives of people by direct action with armed force is linked to the right to life in accordance with Article 2 (2) sentence 1 of the Basic Law Incompatible with the guarantee of human dignity in Article 1. Paragraph 1 of the Basic

Law, in  
aircraft  
Section  
Gazette  
conjur  
conjur

The state should not protect a majority of its citizens by deliberately killing a minority – here the crew and passengers of an airplane. Balancing life against life according to the standard of how many people might be affected on the one hand and how many on the other is not permitted. The state should not kill people because it is less than it hopes to save by killing them.

A relativization of the passengers' right to life cannot be justified by the fact that they are viewed as part of the aircraft as a weapon. Anyone who argues in this way makes them a mere object of state action and robs them of their human quality and dignity.

# The Problem of Full Aggregation

*Death vs Headaches:*

You can save  $X$  from death or a huge number of people from headaches

*Death vs Quadriplegia:*

You can save  $X$  from death or a huge number of people from quadriplegia

[Tomlin, “On limited aggregation”  
Horton, “Always aggregate”  
Halstead, “The numbers always count”]

# Partial Aggregation

All things being equal, we should maximize the sum of strength-weighted, *relevant* complaints

*Competitive Relevance*: A complaint is relevant iff it is sufficiently strong relative to the strongest complaint with which it competes.

*Broad Relevance*: A complaint is relevant iff it is sufficiently strong relative to the strongest complaint under consideration

# Competitive vs Broad Relevance

Case 1: You can save

group A, which contains 1 person facing death and 1 person facing a lost finger,  
or

group B, which contains 1 person facing a lost arm

Case 2: You can save

group A, which contains 1 person facing death,

or

group B, which contains 1 person facing a lost arm and 1 person facing a lost finger

Competitive relevance: Lost finger is relevant in Case 1 but not in 2

Broad relevance: Lost finger is not relevant in either case

# Contradictions: Competitive Relevance

Suppose 10 people losing an arm "equals" one person facing death

Stage 1: You can save either group A (1 person facing death) or group B (10 people facing a lost arm)

Stage 2: 1 person facing a lost finger is added to A, a million people facing a lost finger are added to B



# Contradictions: Broad Relevance

Suppose 4000 people losing a finger outweighs 20 losing arms, and 20 losing arms outweighs one person facing death.

Stage 1: You can save either group A (4000 people facing a lost finger) or group B (20 people facing a lost arm)

Stage 2: 1 person facing death is added to A

Partial aggregation implies a non-transitive utility structure ... which leads to contradictions in the mathematics

But human reasoning is not always aggregative

And not always “rational” in a decision-theoretic sense

*So: What is a human-centric theory of responsibility allocation?*

The reward of a thing well done is to have done it. – Seneca

# What should the self-driving car do?

