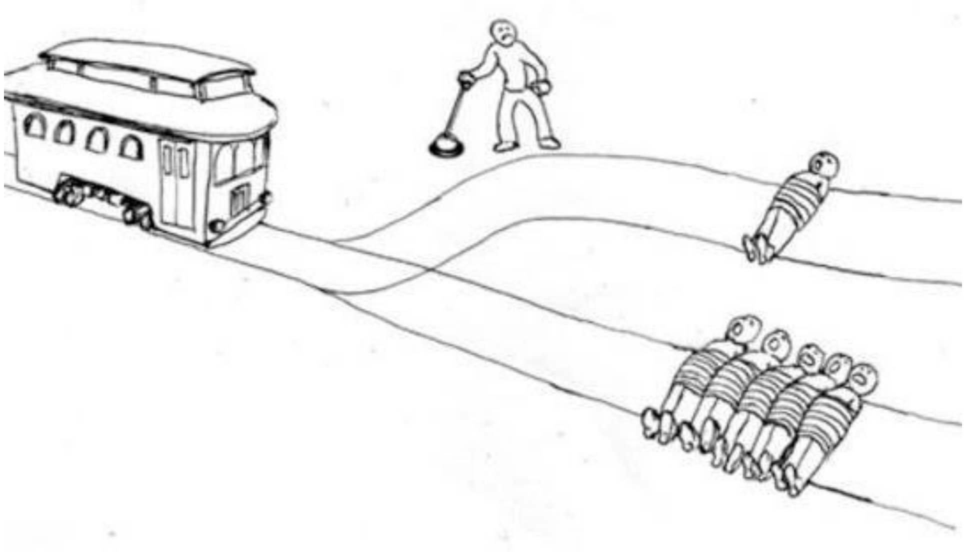
The Trolley Problem and when not to use it

Machine learning and self-driving cars

2021.11.25 | Co-learnium



Australia's National Science Agency

Rules

- Philosophy deliberately stirs up strong thoughts and feelings
- Take note of your emotional reactions, but don't stop there!
- Think about:
  - How do I feel about this?
  - Why do I feel that way?
  - Do I feel differently about situation X compared to situation Y? Why or why not?
  - Is it reasonable to feel that way?
- Thought experiments are to help sort out your feelings, not to apply to the real world

Content warning

- Discussions of:
  - Abortion
  - Euthenasia
  - Harm to/death of babies, children and adults
  - Medical malpractice
  - Fatphobia
  - Ethics and morality
  - AI and ML
- Images of:
  - People stuck in caves

Philippa Foote – The Problem of Abortion and the Doctrine of the Double Effect

- Philosophy has about the same gender balance as physics, ~20% of faculty are women
- Written in the Oxford Review in 1967
- Reproduced in *Virtues and Vices and Other Essays in Moral Philosophy*

Philippa Foote – The Problem of Abortion and the Doctrine of the Double Effect

- The Abortion Act 1967 was passed in the UK on 27 October 1967 and came into effect on 27 April 1968
- Abortion became legal in Great Britain (excluding Northern Ireland) up to 28 weeks' gestation
- Previously (and until 2019 in Northern Ireland) abortion "was illegal unless the doctor acted 'only to save the life of the mother' or if continuing the pregnancy would have resulted in the pregnant woman becoming a 'physical or mental wreck' ". - Wikipedia

The Doctrine of the Double Effect

- There's a difference between the "intended" effect and the "unintended but forseen" effect of an action
  - A doctor administers a large dose of painkilling medicine to a patient in significant pain even though the doctor forsees that the medication will shorten the life of the patient.
  - A doctor administers a large dose of painkillers to kill a patient in significant pain

The Doctrine of the Double Effect

- At the time of Philippa Foote this doctrine was used by those of the Catholic faith to argue against abortion
  - A doctor performs a medically-required hysterectomy on a woman that results in the death of a foetus
  - A doctor administers abortion drugs to kill a foetus

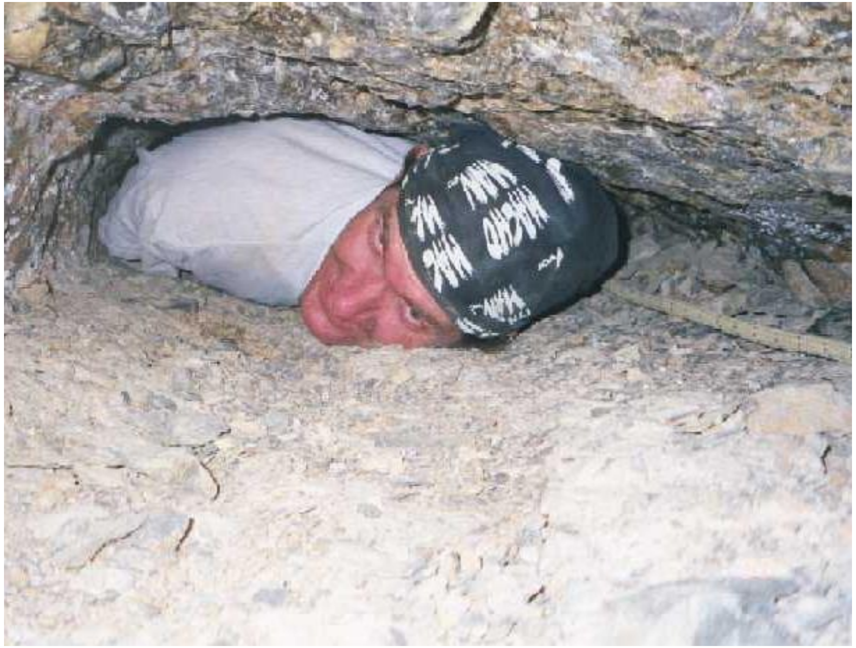
The Doctrine of the Double Effect

- At the time of Philippa Foote this doctrine was used by those of the Catholic faith to argue against abortion
  - A doctor performs a medically-required hysterectomy on a woman that results in the death of a foetus
  - A doctor administers abortion drugs to kill a foetus

  - A doctor needs to decide whether to save the baby or the mother (difficult birth situation)
    - The mother will definitely die in labour but the baby will survive
    - To save the mother the baby must be killed

The Doctrine of the Double Effect

- Fat man in a cave
  - Cavers let a fat man lead them out of a cave
  - Fat man gets stuck, trapping people behind them
  - Floodwaters are rising that will fill the cave
  - Cavers have some dynamite



The Doctrine of the Double Effect

- Fat man in a cave
  - Cavers let a fat man lead them out of a cave
  - Fat man gets stuck, trapping people behind them
  - Floodwaters are rising that will fill the cave
  - Cavers have some dynamite

- Fat man is stuck head out so he would survive the flood
- Fat man is stuck head in so he would also drown

The Doctrine of the Double Effect

- Actual case: merchants selling cooking oil that they knew was poisonous and killing innocent people
- Unemployed gravediggers selling the same oil to create work for themselves

The Doctrine of the Double Effect

- Actual case: merchants selling cooking oil that they knew was poisonous and killing innocent people
  - Selling oil for money
  - Forseen but "unintended" deaths
- Unemployed gravediggers selling the same oil to create work for themselves
  - Selling oil intending deaths
  - Using deaths for money
- (Legally both murderers)

What's the real difference?

- Rioters demand that the culprit for a crime is found and punished or they'll kill 5 people
- Judge doesn't know who real culprit is, so finds an innocent person and has them executed
- **But** we have certain expectations of the law and legal system, so this is a bad example...

The OG trolley problem

- Driver from a runaway tram can choose between two narrow tracks
- One track has five men working on it
- Other track has one man working on it
- He's bound to kill whoever is on the track he chooses

The OG trolley problem

- Most people would say yes to choosing the track with one man on it
- **But** most people are appalled at the idea of framing an innocent man to save five other innocent people from the rioters
  - Same applies if we remove the judge and say that some random person must choose the innocent person to frame
- Foote notes herself that: "Perhaps he might find a foothold on the side of the tunnel and cling on as the vehicle hurtled by. The driver of the tram does not then leap off and brain him with a crowbar."
  - Is that the difference between the cases?

The OG doctor problem

- Drug in short supply
  - Need to give **all** of the drug to one person to save them
  - Or give 1/5 of the drug to five people and save them

The OG doctor problem

- Drug in short supply
  - Need to give **all** of the drug to one person to save them
  - Or give 1/5 of the drug to five people and save them
- Cancer research
  - Kill one person for medical research to find the cure for cancer
  - Save many people with cure

The OG doctor problem

- Drug in short supply
  - Need to give **all** of the drug to one person to save them
  - Or give 1/5 of the drug to five people and save them
- Cancer research
  - Kill one person for medical research to find the cure for cancer
  - Save many people with cure
- Drug made of human
  - Kill one man and make a serum out of his body
  - Save five people with the serum

The Doctrine of the Double Effect

- Puts us in a really really bad position
- Any time some bad guy wants someone to do something, all they need to do is threaten more people
  - If you don't torture this person, I'll torture these five people
  - If you don't kill this person, I'll kill these five people
  - If you don't kill this person, I'll destroy the food supply chain for all of Australia
  - Many many other examples

Alternative theory?

- From *Jurisprudence* by Salmond:

  "A positive right corresponds to a positive duty, and is a right that he on whom the duty lies shall do some positive act on behalf of the person entitled. A negative right corresponds to a negative duty, and is a right that the person bound shall refrain from some act which would operate to the prejudice of the person entitled. The former is a right to be positively benefited; the latter is merely a right not to be harmed."

Trolley problem

- Driver needs to choose between two *negative* duties
  - Injure five people
  - Injure one person
- Therefore there's no conflict
- Can't avoid both, so should choose to do the *least* injury possible

Judge hostage situation

- Has to choose between a *positive* and *negative* duty
  - Positive: rescue five innocent people
  - Negative: kill one person
- Conflict of duties
- Cannot kill one person (perform negative duty)

Doctor problem

- All of drug to one person or 1/5 of drug to five people
  - Two positive duties
  - No conflict – choose to save most people
- Kill one person to save five people
  - Negative duty (kill one) vs positive duty (save five)
  - Conflict – choose not to kill

Conclusion

- What you intend doesn't matter
- Distinction between avoiding injury and bringing aid is more important
- Still has problems:
  - We would all choose to feed our own starving child vs feeding many starving children in another country
    - Both *positive* duties
    - Therefore should choose to do the most good (but we don't)

When to use the trolley problem

- Trying to disentangle ethical and moral problems
- Trying to work out what the actual problem is
- Trying to develop a philosophical theory or idea

When **not** to use the trolley problem

- Self-driving cars?
- Machine learning?
- Other real-world siuations?

When **not** to use the trolley problem

- Self-driving cars?
  - MIT Media Lab designed Moral Machine
  - Experiment to determine the differences in ethics priorities
  - Should self-driving cars prioritise:
    - Human vs pet
    - Passengers vs pedestrian
    - More lives vs fewer
    - Women vs men
    - Young vs old
    - Fit vs sickly
    - Higher status vs lower status (rich vs poor)
    - Law-abiders vs law-benders
    - Take action vs stay on course
  - Various combinations of these scenarios (Three old women vs two dogs)
  - The Moral Machine Experiment by Awad et al 2018

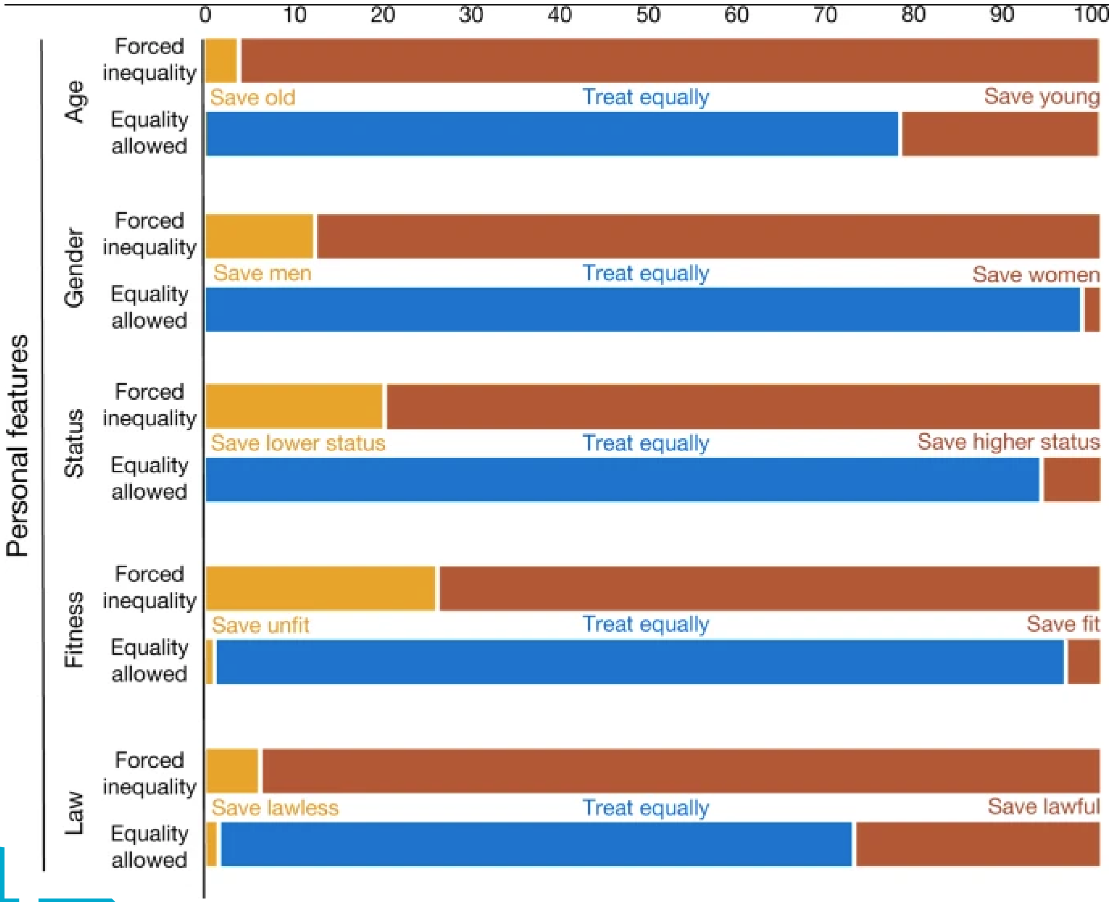
When **not** to use the trolley problem

- Trolley problem results were then used to argue the type of decisions self-driving cars should make (oh no)
  - Real life is not that simple
  - The cars can make other decisions!
  - Depends a lot on the quality of the car, AI, and the information available
  - NOT THE POINT OF THE EXPERIMENT

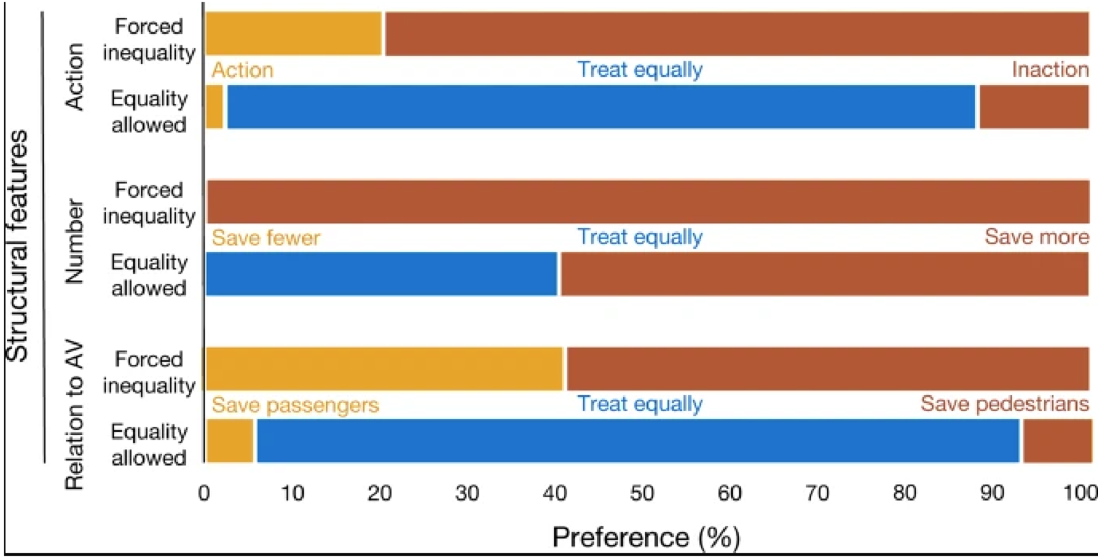
When **not** to use the trolley problem

- *Life and death decisions of autonomous vehicles* by Bigman and Gray 2020
  "Our results challenge this idea, revealing that this apparent preference for inequality is driven by the specific ‘trolley-type’ paradigm used by the MME. Multiple studies with a revised paradigm reveal that people overwhelmingly want autonomous vehicles to treat different human lives equally in life and death situations, ignoring gender, age and status—a preference consistent with a general desire for equality"

When **not** to use the trolley problem



When **not** to use the trolley problem



Australia's National Science Agency

When **not** to use the trolley problem

- Self-driving cars?
  - Better questions than who the car should kill...
    - How can we improve self-driving cars so they don't have to kill anyone?
    - How can we improve the information collected by self-driving cars so they don't have to kill anyone?
    - Are self-driving cars the best way to reduce the road toll?
      - (No, this doesn't mean we should use the trolley problem again)
    - To reduce the road toll do we need only self-driving cars instead of a mix of self-driving and human-driving?

When **not** to use the trolley problem

- Machine learning?
  - Maybe can be used to demonstrate some of the problems with ML applications
    - ML can be used to detect pre-cancerous cells
      - Positive duty - save lives
    - The algorithm was trained on majority white people, so has many false-negatives for black people and people of colour
      - Negative duty – harm/kill people
    - Conflict of positive and negative, so we should choose to prevent harm and we should **not** use the ML algorithm

When **not** to use the trolley problem

- Machine learning?
  - Maybe can be used to demonstrate some of the problems with ML applications
    - ML can be used to detect pre-cancerous cells
      - Positive duty - save lives
    - The algorithm was trained on majority white people, so has many false-negatives for black people and people of colour
      - Negative duty – harm/kill people
    - Conflict of positive and negative, so we should choose to prevent harm and we should **not** use the ML algorithm
  - Assumes that a badly trained algorithm is the only option
  - Better to really think about our ML training **and** to make sure we have a diverse group of people working on ML problems that apply to people

When **not** to use the trolley problem

- Other real-world problems?
  - Probably better to not...

Recommendations

- Philippa Foote – The Problem of Abortion and the Doctrine of the Double Effect (Oxford Review 1967)
- Judith Jarvis Thomson – Killing, Letting Die, and the Trolley Problem (The Monist 1976)
- Awad et al – The Moral Machine Experiment (Nature 2018)
- Bigman and Gray – Life and Death Decisions of Autonomous Vehicles (Nature 2020)