The Philosophy and Ethics of Artificial Intelligence

Dr. Sanjay Modgil

Office hours: 10am – 12pm (Email me to book an appointment)

Overview of Course

- Introduction
- Intelligence
- Consciousness
- Reasoning and Communication
- Ethics and Morality
- Algorithms
- For good or bad ? AI in Medicine, AI in War
- Superintelligence and the Value Loading Problem
- AI and Human Society
- Review/Revision

Ethics and Morality

- Ethical Challenges for Artificial Intelligence
- Philosophical Accounts of Ethics and Morality
- Implementing Moral Agents
- Utilitarianism

Some of the main sources for this lecture:

- Stanford Encyclopedia of Philosophy
- Moral Tribes – Joshua Greene
- Moral Machines : Teaching Robots Right from Wrong: Teaching Robots Right from Wrong. Wendell Wallach

Ethical Challenges for Artificial Intelligence

- Almost everyday stories about risks of AI.
- Why are we more worried now ?

- 1) In the past AI machines deployed in segregated spaces restricted to trained personnel. Now machines - like autonomous cars - may be 'let loose' in non-segregated environments

  2) AI systems in the past used in specific/predictable domains, and with limited *autonomy* (ability to independently make decisions without human intervention/guidance). Hence any possible ethical/legal/societal issues could be handled at development time

  But nowadays, especially because of advances in machine learning:

  i) off the shelf software and hardware available for customization and deployment in unpredictable contexts;
  ii) AI successes suggest greater and greater autonomy (eg autonomous weapons/cars)
  iii) Lack of transparency/explanation as to how ML algorithms make decisions

Ethical Challenges for Artificial Intelligence

Examples

- Autonomous cars. Suppose brake failure. Does the car decide to continue and injure 3 workers on road, or swerve left and injure driver ?
- R&D in autonomous weapons especially in US, UK, China, Russia, France, Israel, and South Korea
- ML algorithms used in deciding insurance premiums, recruitment, mortgage applications, criminal justice system … issues of bias, fairness and transparency
- Sex robots and robots being developed to care for the elderly and disabled.
- Social Media and Internet of Things facilitating, monitoring and analysing human activity. Data privacy.
- Filtering algorithms polarising beliefs and targeted advertising/messaging manipulating and restricting human consumer and political choices
- Medical decision support systems, surgical robots, robots assisting autistic children, image diagnosis
- Superintelligent AI and the Value Loading Problem

What is Ethics, what is Morality ?

Ethics concerns what **we** *ought* to do. What it is 'the good', and how to ensure that our behaviour ensures good outcomes and avoids bad outcomes.

Typically, the terms ethics and morality are used interchangeably

In the context of AI, examples illustrate that the uses of AI, the way we interact with AI, **and the decisions that AI machines make**, have, and will increasingly have, an ethical impact on us

Ethics applied to humans who can **choose** actions that have good or bad effects.

As AIs become more intelligent, and autonomously make decisions, we need to think about how to ensure that AIs are designed so as to behave ethically

Questions and issues when thinking about ethics of AI

1) We are the *users* and *designers* of AI, so there is the issue of

- *our ethical uses of AI* (as has always been the case with new technologies)

- designing AIs to ensure they cannot act other than ethically (as has always been the case with new technologies)

- Unlike past newly designed technologies more autonomous AIs **may have the capacity to act unethically**. How do we design them so that ***they choose ethically good courses of action***
Especially problematic when designers cannot forsee all contexts/ situations in which AGI/SuperAI will act autonomously !

2) AI systems are becoming increasingly embedded within technology so that it is often hard to distinguish which ethical issues are presented by AI itself and which by other features of technology.

3) AI systems are increasingly influencing *our behaviours and social interactions*, so how do we ensure we remain aware of their ethical impact ?

Philosophical Accounts of Ethics and Morality

Key Theories of Ethics/Morality

1) **Deontology** – focuses on duties and rights that are fulfilled and respected by obeying rules, norms and laws. Morality is about finding a good system of rules to guide our behaviour and sticking to these rules

2) **Consequentialism** – focuses on actions that have the 'best' outcomes/ consequences. We should assess the most likely results of our actions and choose the actions with the best results.

3) **Virtue Ethics** – focuses on cultivation of good personality traits/virtues. Persons with these virtues will act in morally good ways. We should work on becoming more honest, compassionate, kind, courageous, etc. As we become more virtuous, we'll be more ethical

Key Theories of Ethics/Morality

Suppose it is obvious that someone in need should be helped.

- A consequentialist will say that the consequences of doing so will maximize well-being.
- A deontologist will say that doing so will be in accordance with a moral rule such as “Treat others as you would want them to treat you”
- A virtue ethicist will say that helping the person would be the act of a charitable or benevolent person.

Deontology

Deontology concerned with choices that are morally forbidden (prohibited), required (obligated), or permitted (recall mention of Deontic Logic in last lecture)

Deontological theories therefore guide and assess our choices of what we ought to do

In contrast to consequentialism, some choices cannot be justified by their effects: no matter how morally good (bad) their consequences, some actions are morally prohibited (obliged)

Two broad approaches; those that focus on the **duties** of the agents that act, to be moral, and those that focus on **rights**

In either case, the focus is on following a set of rules or laws that encode in them moral duties and respect for rights. For example, *you should not kill* encodes both a **duty** and a respect for the **right** to life

Deontology - Examples

Some **deontological** rules governing behaviour are specific in stating what is obligatory and/or forbidden

In *religious moral codes*

- You must not eat beef (a prohibition in Hinduism)

- The Ten Commandments in Judeo-Christian religions, e.g. you must not commit murder, you must not commit adultery….

Secular (non religious) moral codes

- Legal encodings of morality – laws prohibiting murder, theft, tax evasion

- Asimov's laws of Robotics

Deontology - Examples

Some **deontological** rules governing behaviour are more abstract – they provide general principles that need to be 'appropriately' applied in any given situation

E.g. *The Golden Rule*

> Treat others as you would like others to treat you

E.g. Immanuel Kant's (1785) *Categorical (*unconditional*) Imperative* (command)

> Act only according to that rule by which you can at the same time will that it should become a universal law.

The categorical imperative (CI) commands that every rule you act on must be such that you are willing to make it the case that everyone always acts according to the same rule when in a similar situation.

Kant's Categorical Imperative

Eg suppose you are considering whether to act according to the rule:

if I want something it is permitted to lie to get that something

I would have to be willing that everyone lied to get what they wanted - but then no one would believe you (in fact no one believe anyone), the lie would not work and you would not get what you wanted. So, if you *willed* lying to become a universal law, **you would not achieve your goal**. Therefore according to the CI it is not permitted to lie. It is not permitted because the only way to lie is to make an exception for yourself.

Kant believed the above is a purely *rational recipe for how to act* - **if you don't act according to the CI your own goals would be blocked**. He believed it was equivalent to the more explicitly moral:

> Act so that you treat humanity (yourself and others) always at the same time as an *end* and never simply as a *means*.

E.g., Slavery would be wrong because slaves would be used as a means to achieve the goals (ends) of the slave owner (e.g., cultivation of crops)

Critiques of Deontology

1) Too general/vague and so doesn't always guide action in a specific situation

2) Inflexible - *absolute* prohibitions on certain actions (including murder, theft, and lying). What about exceptions ? For example, it is permissible to lie when doing so will save a life, or it is permissible to kill in self defence. **But** these rules could in principle be updated to include exceptions and which then would be rules we would want to make everyone follow.

3) Deontologists usually fail to specify which principles should take priority when rights and duties conflict, so that Deontology cannot offer complete moral guidance

4) Applications of rules mean that overall **consequences** of actions may be harmful

- a sadist is a masochist who follows the Golden Rule

- *CI* only allows killing as a ***side effect*** of doing something for the greater good and not as a ***means*** to achieve the greater good – e.g. prohibited from killing 1 person to save 100 or 1000 or ….

Virtue Ethics

Virtue Ethics describes character of a moral agent as a driving force for ethical behaviour and originates in the ethical theories of Socrates and Aristotle

Virtue Ethics not so much concerned with the effects/consequences of actions (consequentialism) or fulfilling duties/respecting rights by following rules (deontology)

> Virtue Ethics says that morally good actions flow from the cultivation of good character, which consists in the realisation of specific virtues

e.g. courage, truthfulness, modesty, generosity, wisdom, justice …

What is good is **learnt** from particular actions, from making connections between actions and their consequences

Humans learn what is good through intuition, induction and experience. Eg by asking good people about the good, one's sense of the overall goal of the good comes into focus, and the ideal person acquires practical wisdom and moral excellence

Critiques of Virtue Ethics

- different people, cultures and societies often have different opinions on what constitutes a virtue, perhaps there is no one objectively right list.

- Does not provide guidance on what sorts of actions are morally permitted and which ones are not, but rather on what sort of qualities someone ought to cultivate in order to become a good person.

Consequentialism

Consequentialists say that actions should be morally assessed only by the states of affairs they bring about. They specify states of affairs that are good, and then say that whatever actions increase the Good are morally right.

The consequences of one's actions are the ultimate basis for any judgment about the rightness or wrongness of those actions.

Unlike deontology, no notions of rights, duties, rules. No rules to say that some actions are prohibited/obliged. All that matters is the consequences. Eg. one might lie in a court of law, to save an innocent person's life, even though it is illegal to lie under oath.

Consequentialists can and do differ widely in terms of specifying **The Good**. One of the most popular and well known versions defines the good is in terms of 'happiness'

Utilitarianism

Late 18th- and 19th-century English philosophers Jeremy Bentham and John Stuart Mill proposed that actions are right if they tend to promote happiness and wrong if they tend to reduce happiness—not just the happiness of the performer of the action but also that of everyone affected by the action.

Bentham and Mills argued against slavery on utilitarian grounds and not by appealing to rights !

Utilitarianism = *Impartial Maximisation of Happiness*

- *Impartial* (related to the *Golden Rule*): **everyone's happiness counts equally**

Suppose you had a choice between two actions A1 and A2. A1 increases the *average* happiness more than A2, so choose A1

We will look at utilitarianism and its critiques in more detail later

Moral Relativism

Moral relativism argument:

*Premises***:**

1) morality is simply the expression of ethical or **value** judgements we make in a given society

2) Other societies have their own judgements, we have ours

Conclusion:

Therefore we **should** not judge other cultures

But

The conclusion itself a 'value judgement' – that is to say it is making a judgement about what one should/should not do – and it is announced as if it were some universal truth – no matter what culture you are in, you should not judge other cultures

Contradiction !

Implementing Moral Agents

Moor's Categorisation of Moral Agents

Ethical Impact Agents: Any machine that can be evaluated for its ethical impact if the agent's operations increase or decrease good in the world (e.g. replacement of camel jockeys - young boys – with robot jockeys in Gulf states)

Implicit Ethical Agents: Machines **designed** to not have negative ethical effects – behaviours constrained by designers following ethical principles. Programmer anticipates possible courses of action and provide rules that lead to the desired outcome in the range of circumstances in which the agent is to be deployed (e.g. warning devices to alert pilots when another plane is approaching on a collision path) – *no explicit representation of ethical principles*

Explicit Ethical Agents: Machines that can reason what is the best action in ethical dilemmas, and novel situations that may not have been anticipated by designers, using ethical principles – *explicit representations of ethics e.g. in the form of deontic logic rules (do you think the absence of explicit representations of ethics is a necessary or sufficient condition for stating that an agent is not an explicit ethical agent)*

As agents become more intelligent (*ability to achieve goals in a wide range of environments*) and autonomous (independent from human guidance) they may need to be *explicit moral agents*.

Moor's Categorisation of Moral Agents

Full Ethical Agents: Machines that can be said to have '**moral agency**' and are able to justify their moral judgements.

In philosophy and law *moral agency* equates with *moral responsibility* :

if one **understands** the ethical impacts of one's actions and **freely chooses** action that is ethically wrong, then one is morally responsible in the eyes of the law, and so should be held accountable for one's actions

A question for you to consider. Are these categories mutually exclusive ?

Implementing Artificial Moral Agents

Three approaches to designing and implementing ethical agents (artificial moral agents - AMAs)

- Top Down

An existing moral theory (ranging from abstract to specific) is chosen and encoded in the agent which applies the theory in concrete situations so as to act ethically – typically logic based and deontic (GOFAI)

- Bottom Up

Agent explores course of action and learns and is rewarded for behaviour that is morally praise-worthy – in other words *reinforcement learning* - ethical values implicit in agents' activities, rather than explicitly articulated.

- Hybrid

Uses elements of both top down and bottom up

I will focus on top down and in a later lecture discuss bottom up (machine learning) approaches

Top Down Implementations of Deontic AMAs

Deontic Logic Encodings of Ethics in Agents

Deontic Logics well studied and supported by a variety of programming frameworks

Deontic modalities in Deontic Logic

Oq = *q* is obligatory
Pq = *q* is permitted
Fq = *q* is forbidden/prohibited

Autonomous Car in UK Example

F drive_on_right
O protect_driver

But what if car in a two way tunnel and collision anticipated as highly likely given braking distance, and calculated risk of injury to driver is high

Only way to avoid collision is to swerve car and drive on right ? There is then a conflict. Hence prioritise obligations:

Top Down Implementations of Deontic Agents

collision_imminent $\wedge$ *risk_estimation(injury, left, X) > 0* $\longrightarrow$
O protect_driver > F drive_on_right

but only if risk of injury due to collision with oncoming traffic on right < risk of injury due to collision if stay on left

collision_imminent $\wedge$ *risk_estimation(injury, left, X)* $\wedge$ *risk_estimation(injury, right Y) and Y < X* $\longrightarrow$ *O protect_driver > F drive_on_right*

but what if car not in tunnel and so can swerve off the road to the left, or what if non-zero risk of injury to occupants of car in oncoming traffic on right, and only 1 such occupant or 2 occupants one of which is a young child …..

As we discussed earlier in this lecture and in *Reasoning and Communication* the only way to guarantee the "correct" conclusions is to explicitly list all possible circumstances and exceptions in which rules apply and when to prioritise one rule over another. This is even more critical when reasoning about ethics !

Top Down Implementations of Deontic Agents

So rule based (deontic logic) specifications of correct behavior ok if agent deployed in very restricted environment and rules encode all forseeable situations

But for autonomous agents in complex and unpredictable environments and hence environments in which one cannot predict all situations that may arise:

deontic specifications of ethical behavior need to be more general

and so the challenge is:

How to interpret general principles in any given specific situation so as to ensure correct ethical choices. Specifically:

- How to interpret in a commonsense way the intentions underlying principles ?
- How to prioritise in a commonsense way conflicts amongst ethical rules ?
- How to avoid deadlock when conflicts arise from rules with equal priority ?
- How to avoid unforeseen/unintended harmful consequences of following rules (*a sadist is a masochist who follows the golden rule*) ?

Asimov's Four Laws of Robotics

Issac Asimov's novels and short stories highlighted these problems

In order of priority

0. A robot may not injure humanity, or, by inaction, allow humanity to come to harm (a generalisation of law 1)
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 the 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 Laws

- What does it mean to harm a human ? A literal minded robot following Law 0 might interrupt a surgeon about to cut a patient

Again, unless context fully described and understood (does not count as harm if goal to make patient healthy and patient gives consent and patient anaesthetised ..)

Implementing Kant's Categorical Imperative

- How would an AMA implement the CI ?

- *Ideally* a powerful AMA, when considering an action according to rules that are either programmed from the outset or learnt over time, runs a simulation to determine whether its goal would be blocked if all other agents were to operate according to the same rule.

- Again requires a great deal of understanding about context, causal reasoning, human psychology …

Virtue Ethics and Implementing Bottom Up AMAs

- The bottom up machine learning approach (i.e., reinforcement learning) can be understood as equating with virtue ethics

- Both emphasise development of ethical behaviours through training and learning, through perceiving how actions lead to good ends without explicit symbolic rules governing behaviours

- A ML agent will seek to maximise rewards through optimising utility functions – in the lecture on superintelligence we will review some of the issues and challenges that arise when relying on these bottom up machine learning approaches.

Utilitarianism

Utilitarianism and Happiness

Utilitarianism = *Impartial Maximisation of Happiness*

Consequentialism – ultimate goal is to make things go as well as possible.

Utilitarianism = consequentialism + what really matters is *happiness*

- **Why** happiness and **what** is happiness ?

Why:

1) True for everyone and so a basis for a *universal* morality - what ultimately matters is the positive quality of our experience (happiness) and this is true for everyone

2) Hard to think of values worth valuing (whichever the society/culture) that are not ultimately linked to happiness: Values associated with personal relationships (family, friends, love …), personal virtues (honesty, wisdom …), noble pursuits (truth, art, sport …), and good governance (freedom, justice …)

Subtract from all the things we value, *their positive impact on experience*, and their value is lost. So happiness is:

Utilitarianism and Happiness

What:

1) The universal underlying value

2) The **subjective** experience of well being

3) Think 'counter-factually' – suppose something we value was absent. Would the absence reduce our happiness ? Then that thing makes us happy

Can we in principle measure happiness ?

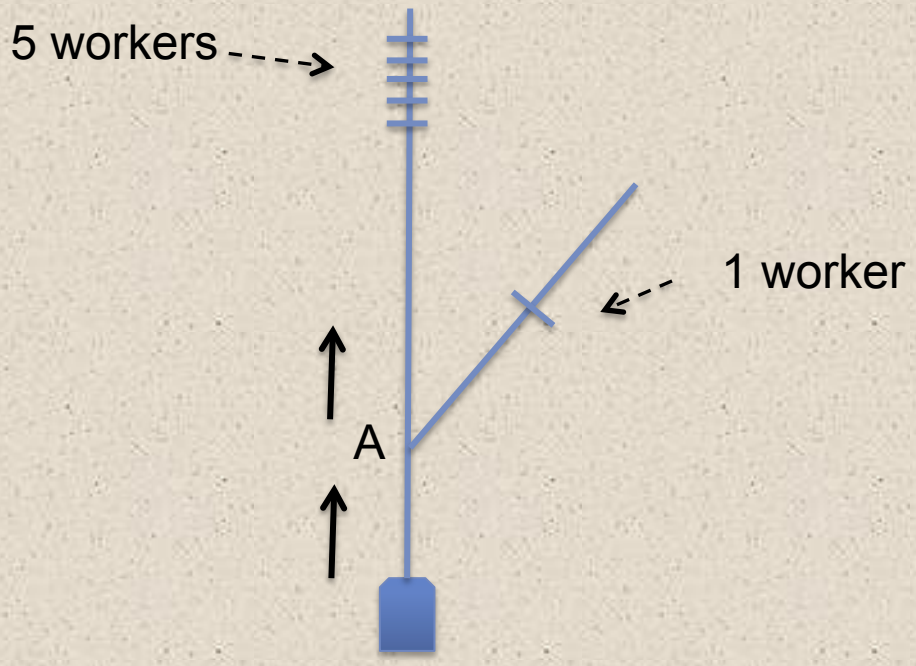
- Brain science has done a lot to identify neural correlates of happiness.
- Asking people how happy they are is *usually* a good way of measuring
- In 'economics' lots of research into how to measures happiness (*gross national happiness* instead of *gross domestic product*)

As an aside listen to Bobby Kennedy's excoriating attack on the notion of GDP as a measure of a country's worth: https://www.youtube.com/watch?v=3FAmr1Ia6w0

Trolley-ology

Let's explore utilitarianism with thought experiments that involve a trolley on a track hurtling towards 5 workers who are tied down and will be killed by the trolley (in all these experiments you know nothing about the workmen, e.g. their ages or whether some workmen are better people than others)

The Switch Dilemma

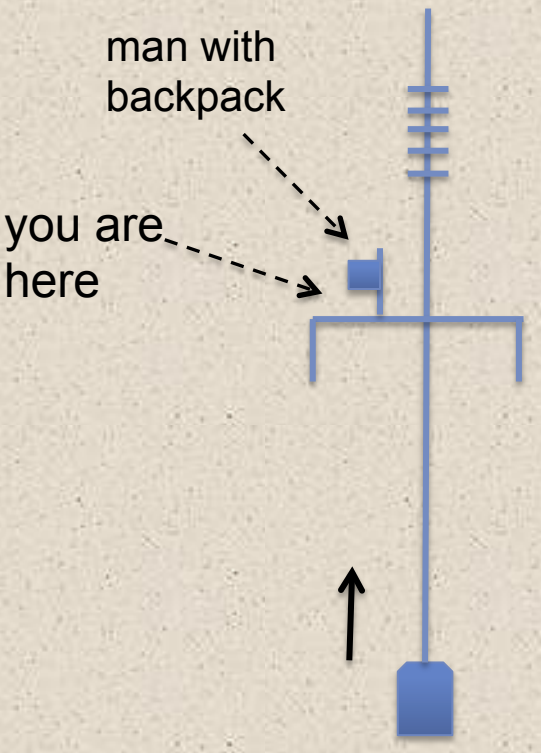


At point A you can steer the trolley so that it takes the side track on the right and kills the one worker instead of the five

What would you do ?

Trolley-ology

The Footbridge Dilemma 1

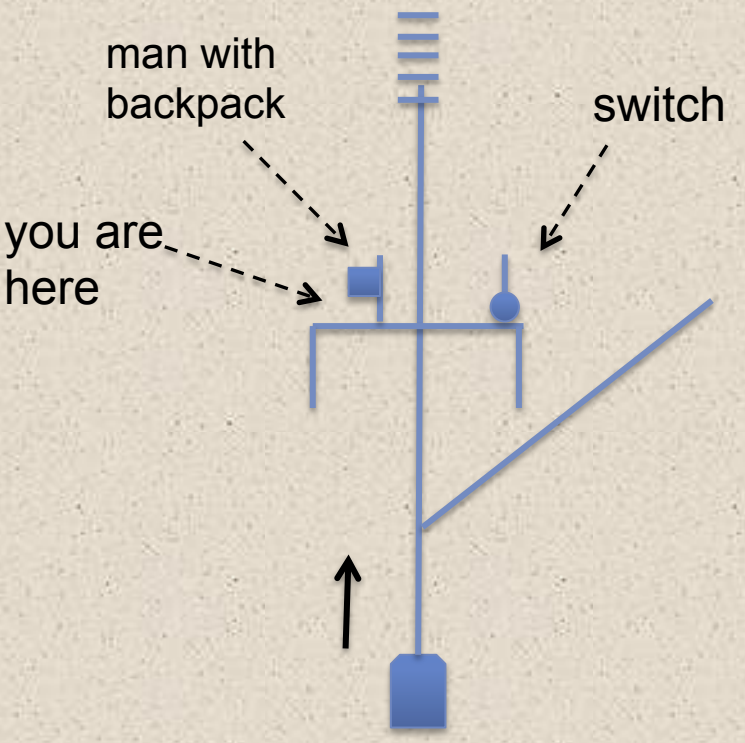


You are on a footbridge and can push a man with a backpack of the bridge so *guaranteeing* that the trolley will be stopped. The 5 workers are saved, but the man with backpack dies instead – he is used as a 'trolley stopper'

What would you do ?

Trolley-ology

The Footbridge + Switch Dilemma



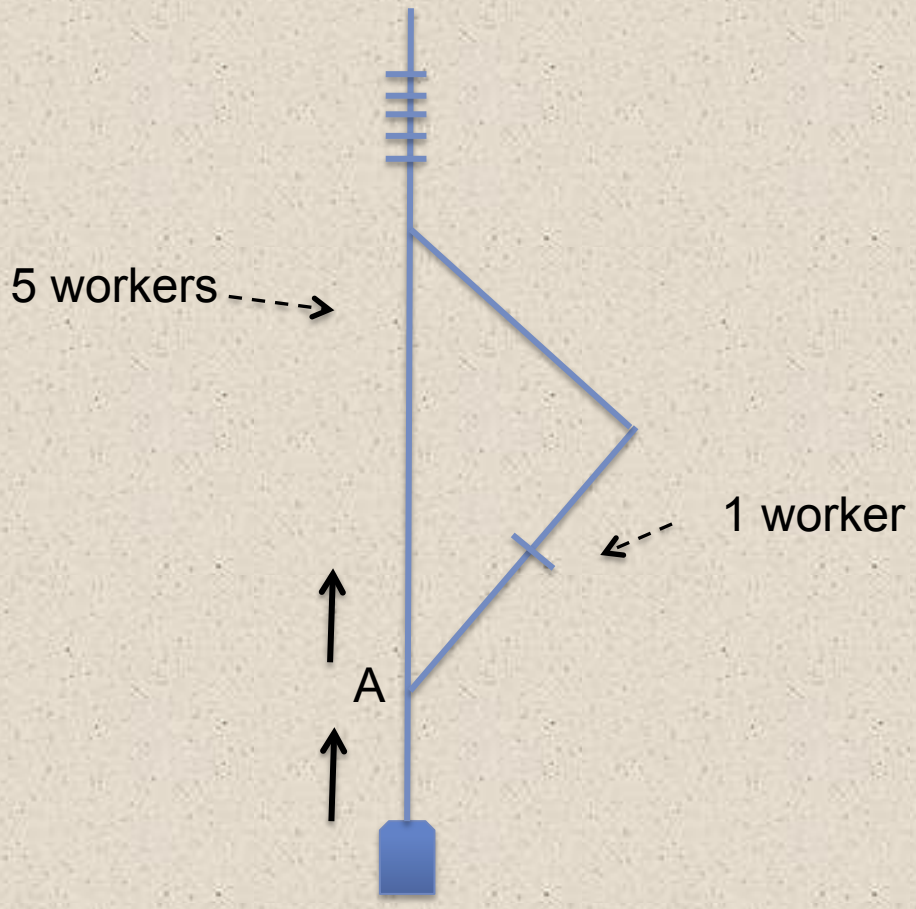
You are on a footbridge and need to run very quickly (no time even to warn man with backpack) to the end of bridge to pull switch so that trolley takes track on right and saves the 5.

But you know that it is impossible to avoid colliding with the man with backpack and knock him of bridge so that he dies from the high fall

What would you do ?

Trolley-ology

The Switch Dilemma 2

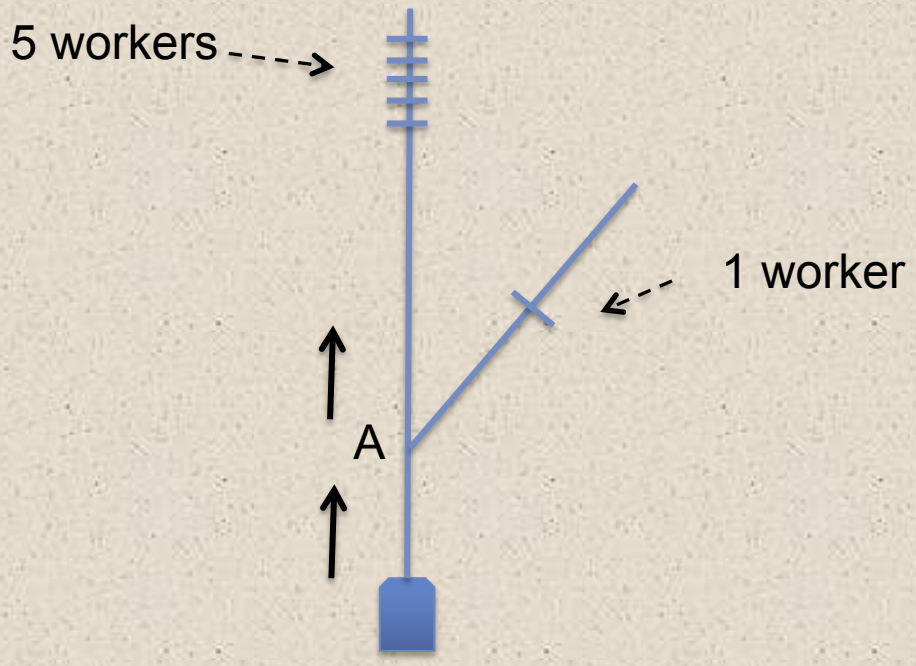


At point A you can steer the trolley so that it takes the side track on the right and kills the one worker instead of the five. But note that this time, if the worker were not on the side track, the trolley would return to the main track and kill the five

What would you do ?

Trolley-ology

The Switch Dilemma



What would you do ?

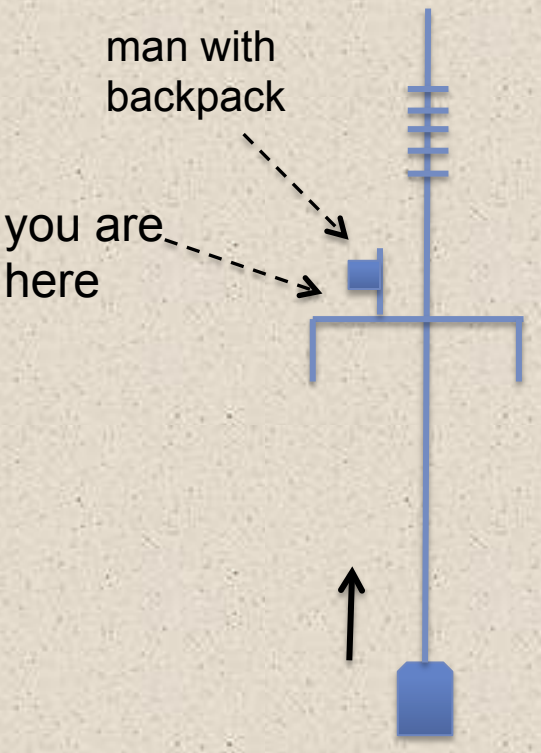
87 % would pull switch

Physical Force And Means to End

The Footbridge Dilemma

What would you do ?

31 % would push man



But the utilitarian calculation is the same !

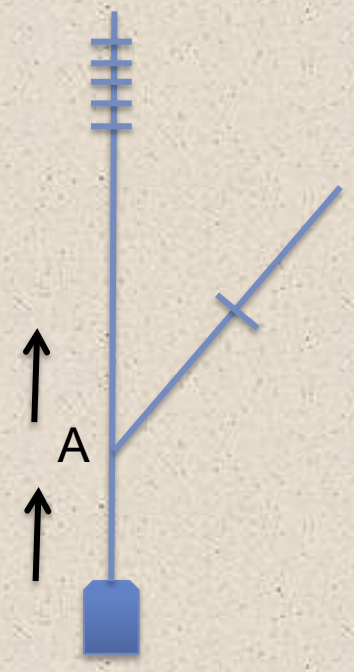
What's the difference ?

1) In this case you're actively pushing the man and our emotional resistance to using physical force interferes with rational calculation

2) The man is used as a trolley stopper – as a *means to an end*. If you did not push the man of the bridge, you would not achieve your goal to save the five.

No Physical Force and Side Effect

The Switch Dilemma



What would you do ?

87 % would pull switch

But in this case:

1) No physical force

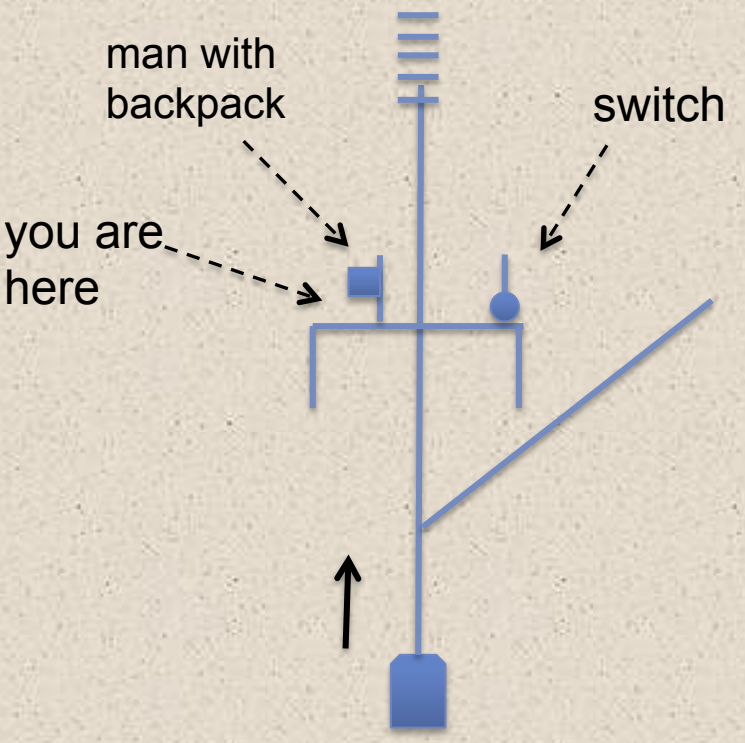
2) The killed worker is not used as a means to an end – as a trolley stopper – his death is a *side effect*.
Think *counter-factually*: if the 1 worker were not tied down on the right hand track, you would still achieve your goal of saving the five. Indeed it would be great if there was no one tied down.
So killing the one is not **causally** important to saving the five – killing the one is not a means to achieving the end (goal) of saving the five

Physical Force and Side Effect

The Footbridge Dilemma 2

What would you do ?

81 % would run to pull switch



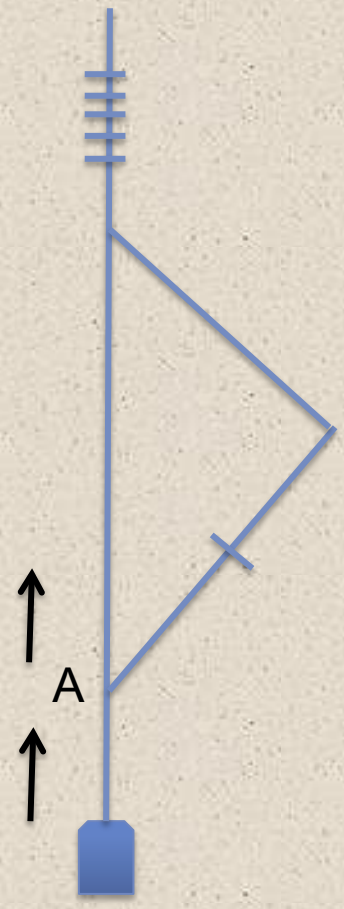
1) Physical force

2) The death of the man is a side-effect and is not used as a means to an end – as a trolley stopper – his death is a side effect

Again, thing counter-factually
(if the man were not there then great !)

No Physical Force and Means to End

The Switch Dilemma 3



What would you do ?

81 % would steer the trolley right

1) No physical force

2) The worker is used as a means to an end – as a trolley stopper – if the man were not there the five would be killed. So killing the one worker is required to achieve the goal of saving the five

Trolley-ology - Conclusions

It is the ***combination*** of physical force and means to an end that is important (that interferes with utilitarian calculation) Why ?

1) *We have evolved to have an emotional reluctance* to violence as violence disrupts cooperation in societies as well as our own survival

2) *Instinctive* distinction between *means to ends* and *side effects* interferes with our rational utilitarian calculations. Why this instinct ?

   Because when we think about plans of action our emotional alarm system responds to harms that we anticipate, and means that are causally necessary for achieving goals are more clearly represented cognitively, unlike side-effects of our actions

In international law it is illegal to *intentionally* bomb civilians to lower the enemies moral (they are used as *means*) but legal to bomb munitions factory knowing full well that civilians will die (side effect / 'collateral damage')

But why should there be any moral distinction between the two ?

Trolley-ology - Conclusions

We should apply utilitarian principles to make bombing with *forseeable* collateral damage illegal, since *distinction between means to an end and side-effect are morally irrelevant*

But instead we allow our emotions/moral intuitions to influence our ethical decisions and arguably make the wrong decisions. **Our emotions/moral intuitions are generally sensible but not always correct**

Trolley-ology - Conclusions

Overall Conclusion

Some people use trolley-ology thought experiments to show that there is something wrong with utilitarianism. In some situations, people apply utilitarian calculations, in others they don't.

But what people (emotionally) think is the right thing to do doesn't *necessarily* make it the right thing to do.

Future AIs will not have their utilitarian reasoning biased by evolutionary acquired emotional instincts.

Other (supposed) arguments against Utilitarianism

Utilitarianism is not demanding enough – sometimes it leads to intuitively unethical choices

Suppose five people, all who will die if they don't receive (different) organ transplants. Utilitarianism says that it would be ok to 'harvest' the organs from another person who would then die, and use these organs to save the five

Do you think Utilitarianism says that this would be ok ?

What kind of society would be a better/happier society – what kind of society would impartially maximise happiness ?

One in which this was a legally allowed practice or one or one in which this was not a legally allowed practice ?

Other (supposed) arguments against Utilitarianism

Utilitarianism is too demanding – it requires making sacrifices that are unreasonable

Suppose you saw a drowning child. Would you jump in and save her if at the same time you ruined your $500 suit.

Of course you would !

So, if you are a utilitarian, instead of spending $500 on a suit, you should donate to save the lives of children in poverty

But **pragmatic utilitarianism** says: Being a perfect utilitarian is anti-utilitarian since it is incompatible with the way our brains are designed.

Practical application of utilitarianism needs to be compatible with your good life => you will be a good role model that others will emulate (*but not if you make impractical self-sacrifices*) and so lead to the greater good !

Homo utilitus is an ideal to aim at – recalibrating our moral stance while acknowledging that we are human.

Back to the Debate on Abortion

Ultimately Pro-choicers appeal to the (deontic) **right of the woman to choose**

Ultimately Pro-lifers appeal to the (deontic) **right of the foetus to life**

In both cases, different moral emotions/instincts are effectively expressed in terms of **rights**

But what justifications are there for these rights ?

Unlike rights, consequentialist/utilitarian appeal to greater good ultimately appeals to evidence. Whether policy will increase or decrease happiness can be answered **empirically** (by examining the data/evidence)

Of course, when moral matters have truly been settled (e.g. slavery) makes sense to talk about rights - they express our firmest moral commitments. But when faced with complex and controversial moral dilemmas, utilitarianism may be the answer

Back to the Debate on Abortion

In the case of the abortion debate:

Poverty reduced when women can abort. No illegal life threatening backstreet abortions. Decrease crime. Allow sex for pleasure

Perhaps outweighed by happiness of future lives especially with availability of adoption. But then prolifers should be against contraception and abstinence – since both will decrease lives. Too much to ask of real-world non-heroic people

Back to Implementations of AMAs

Utilitarian AMAs face the same computational problems of utility maximizing (AIXI/Russel Norvig) intelligent agents

Computational intractability of causal reasoning and estimating effects of actions on all beings that are capable of experience happiness (all conscious beings).

On the other hand

- we could satisfy ourselves with utilitarian calculations that are good enough under resource bounds

- weather forecasting suffers same problem, but this hasn't stopped meteorologists making more accurate predictions - they average across the predictions of several computer models - utility forecasters might do something similar.

Combining Theories in Implementations of AMAs

Perhaps AMAs will need to combine all three kinds of ethical theory

Machine learning will reward AMAs for ethically good behaviours (virtue ethics) and equip them to do the right thing in situations that have been experienced before and where there is little controversy as to what is the right thing to do.

More controversial and novel (and so no experience to draw on) ethical dllemmas will be decided by utilitarian calculations involving the superior (compared to human) data/evidence collection and causal reasoning of the AMAs, combined with humans providing accounts of how actions impacts on human experience.

Settled controversies then encoded as rules that guide long term planning and behaviour (combining deontology and utilitarianism = *rule utilitarianism*)