Lesson 2

Ethics / Morality

What is morality/ethics

- In deciding what to do, or in evaluating what others do:
    - We can take our individual perspective, focusing on our particular interests (**self-interest**)
    - We can be motivated by the belief that an action is right, regardless of how it affect our interest (**morality**)
        - we take a sort of impartial perspective
- **Positive (conventional) morality**: the moral rules and principles that are accepted in a society
    - Can there be bad positive morality?
        - For example in certain societies it may be considered right if a woman is subordinated to her husband and this is seen as an ethical requirement. From our perspective we can consider that this thing is not so good. Assuming that we are working as managers of a company and in our business domain it is believed that as a manager i have to maximize the share of the value of a company regardless on the impact that a company activity has on the environment, on the workers and so on. There are various circumstances that have ethical norms that are followed in the society (**positive/conventional morality**) and then there is also what is called **critical morality** (people may have the view that the positive morality followed by the morality is correct or wrong and may criticize it on the base of an ethical approach that they believe could be better).
- **Critical morality**
    - The morality that is correct, rational, just (maybe since considers all individual and social interests at stake giving each one the due significance (harms to other, impacts on environment, etc.)

        - Critical approach meant to identify what is correct even if it is not shared /accepted among the society

- We can criticise positive morality based on our critical morality: we may be right or wrong (e.g., feminist critiques against patriarchy, nazi criticism against liberalism)

Ethics vs metaethics

- Normative ethics is concerned with determining what is morally required, how one ought to behave
- Metaethics is concerned with is the study of the nature, scope, and meaning of moral judgement
    - **Metaethics** concern with the nature of judgement which cannot be true or false because is based on a feeling approval
    - Can ethical judgments be true or false?
        - What is the difference between
            - I prefer vegetables to meat
            - Vegetables are better than meat
            - I ought to eat more vegetables to me more healthy —> Kant called this statement '**hypothetical imperative**' or **'prudential statement'** —> we do not have an obligation of being more healthy —> is not an ethical obligation.
                - It is true that if you eat more vegetables you become more healthy and this is also a preference for pursuing the goal to be more healthy by eating vegetables. We don't have an obligation to be more healthy, it's something that simply we may do if we want to be healthy. If we want to be healthy, eating vegetables could be a good thing.
            - We ought to become vegetarians —> this statement looks like an **ethical statement** —> it presents as a claim that is valid for everybody, regardless they have they desire to be more healthy. It has the obligation to become vegetarian. Even if something does not care of the planet, you can say that he has to become vegetarian because this compete to this purpose which is bounding from everybody.

            - What facts make this ethical statement true? Is there anything in the world that makes true one of the others prepositions?
- Do they correspond to some facts in the world? What facts an ethical statement true?
    - What facts make it true that we ought to become vegetarian? Or that we ought not to harm others?
    - There are different approeaches to metaethics
        - **cognitivist:** is possible to know to our rationality what should or not should do
        - **non-cognitivist** —> emotivists
        - realist vs non realist approach —> according to realists there are facts in the world that make ethical statements true.
- Does ethic pertain to rationality of or to feelings?
    - **David Hume**: is not contrary to reason to prefer the destruction of the whole world to the scratching of my **finger**. Morality is a matter of having the right feelings
        - "ethics is not the contrary of reason, is a matter of having the right feelings/sentiments"
    - **Emmauel Kant**: we can know what is moral through our reason
    - **David Ross**: we can know what is more through our intuition

Notes about Law vs Ethics

Law concerns those kind of norms that have been delivered through an institutional process by the legislators or by judges or by customs, and laws are enforced coercively, that is if one violate the law then there should be a judicial proceeding that ends up in various outcomes.

On the contrary if we remain in the domain of ethics, if you violate an ethical rule, it is a matter for your consciousness. If this ethical rule is also socially shared you can have the negative opinion of your fellows. If this rule is also a legal rule (for example the commission of omicide) then you end up also in the legal consequences.

It could be that the laws enforce the development of positive ethics and critical ethics may be different from law. But it could also be that there is a norm on which agree but we do not want that it becoemes legally enforced (for instance we should be kind to our friends -ethical requirement- but we would not want that the obligation to be kind becomes a legally enforced norm that regulates that when someone is not kind then it should be punished).

So there is a partially overlap between ethics and law on both sides: on one side not all the laws concern ethical requirements neither in positive nor in critical ethics and is not the case that all ethical requirements are also legal requirements.

This also apply to AI: consider for example that we have an AI system that makes inferences about people and this system has been training on a lot of data and the system makes an inference concerning me and according to this inference I am a bad lander. This is a personal information of me that is inside the system. Should I tell to X that is considered as a bad lander/borrower for my system? Is this a legal requriement? Maybe yes, maybe no because we have the GDPR that requires data subjects to be informed when their data are processed and X knew that his data were processed by the bank, but he does not know that is considered as a bad borrower. Is this inference done by the system to be considered as a personal info present in the system? Maybe yes, maybe no.

I have no legal requirement to say that X is a bad borrower. But do I have an ethical duty that the system should be transparent? Maybe yes.

So, if I conclude that I have this ethical duty, there will be a duty that is not a legal duty.

Morality and disagreement

- Morality is a place for widespread disagreement
    - Abortion
    - Migration
    - Capital punishment
    - Humanitarian wars
    - …
- But there is something on which we may agree? How can society arrive to a single opinion and write a documents such as the document of trustworthy AI?

    - It is wrong to kill innocent people?
    - It is (usually) wrong to lie?
        - Is it possible that lie is good in some occasions or it's always a bad thing? Can you imagine situations in which say something false is good?
    - It is (usually) wrong to harm people?

Pro-tanto and all-things-considered moral judgement

- Many moral prescription are defeasible. They state general propositions that are susceptible of exception.
    - We should not lie
    - What if a lie would save a person's life?
- Do we want a robotic agent to take its duties as defeasible?
- An act is a ***prima facie* duty** when there is a moral reason in favor of doing the act, but one that can be outweighed by other (moral) reasons.
    - Many moral prescriptions are defeasible that can be subjected to exceptions —> we should not lie, but what if lie save a person's life?
    - When we develop robotic agents, what do we want from those agents? Do we want robots that are able to make exceptions or do we want agents that always comply with the norma that have been provided to them?
- David Ross: "If I have promised to meet a friend at a particular time for some trivial purpose, I should certainly think myself justified in breaking my engagement if by doing so I could prevent a serious accident or bring relief to the victims of one.
    - David Ross anticipated the feasible reason in AI

Morality and other normative systems

There are various normative systems that can we follow

- **Law**
    - Does positive or critical morality include all laws enforced by the state? Does it include only such laws?
        - There is an overlap but not complete: neither the law is included in morality, neither the morality in law —> there are legal laws not moral

and moral statements that are not legal

- **Religion**
    - Does critical morality include all and only what has been commanded by God
        - Does God commands something because is moral (bounded by morality) or is the creator of the morality?
        - Did God command something because it was moral, or di anything become moral for having been commanded by God (**rationalism** vs **voluntarism**). What about Abraham and Isac.
        - Are atheists necessarily immoral or amoral? Is an atheistic society necessarily more immoral than a religious society?
- **Tradition**
- **Self interest**:
    - may morality and self interest collapse: should we do all and only what fits our personal interest (Gige's ring)
        - morality is just replication of what is interest

We move into the analysis of **moral theories**.

Consequentialism

This is a very common moral approach.

The concept of consequentialism

Idea: we should judge actions by considering their outcomes.

- An action is morally required
    - iff it delivers that best outcome, relative to its alternative
    - Iff its good outcomes outweight its negative outcome to the largest extent
        - Morality from the pov of consequentialism seems like an **optimization problem**
    - Iff it produces the highest utility?
        - no alternative that produces the highest outcome

- Morality as an optimisation problem!
- Various kinds of consequentialism
    - What are the good and bad things to be maximised?
        - for utilitarians, we need to maximize the satisfaction of people.
    - How many there are?
    - How much each of them matters?
    - Can we construct a single utility function that combines gains and losses over multiple valuable goals?

The reference approach: Utilitarianism

Utilitarianism = key approach to consequentialism

- Jeremy **Bentham**, (also Cesare Beccaria)
- John Stuart **Mill**. From Utilitarianism (1861). **Principle of utility**:
    - Actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness. By happiness is intended pleasure, and the absence of pain; by unhappiness, pain, and the privation of pleasure
- **Utility**: Happiness or satisfaction of desires/interests
- **Utilitarianism is not egoism**
    - egoist thinks only about its own interest. Utilitarians take into account the impact of actions on everybody. Action that brings more happiness than pain and for everybody
    - Example AI recommending system: We have to see whether a recommending system a positive or negative impact. —> **We have to forecast what would be the impact of the choice on the product on the people well being**. Positive impact if the recommending system suggests a product that we like, the negative impact would that by relying on suggestion the recommendation can move us into something that does not make us happy, so in the end the outcome may be negative.
    - We have to **choose the action that, among those available, provides the better payoff.**
    - **Choose the action that provides the better payoff.**

    - **The utility of everybody has to be taken into account equally**

Advantages of utilitarianism

- Conceptually simple
- Egalitarian (everybody’s utility counts in the same way)
- Fits with some basic intuitions (making people happy is good, making them suffer is bad)
- In many case it is workable (what does it say about hunger, should we donate?)
    - on hunger an utilitarian would agree because we suffer a bit by giving our money away, but on the other hand the happiness is largely outweight by the happiness that other people would get from our money.

Two versions of utilitarianism

1. **Act utilitarianism**
    - Do the action that maximises utility
    - Do the optifimic action
2. **Rule utilitarianism**
    - Follow the rule the consistent application of which maximises utility
    - Follow the optifimic rule
    - Rule utilitarianism says: we should not think about actions, we should think about rules and with the application of the rules we can maximize the utility.
        - For example 'do not harm people' is a rule, so we just think to the rule knowing the outcome (so also without acting). While 'act utilitarianism' compute the outcome only after performing some action.
        - Rule utilitarianism is more difficult tan act utilitarianism because you need to have in mind all the possible outcomes.

- Is AI utilitarian
    - What utility function would be utilitaristic?
        - Utility function is a function that quantify a goal that a system needs to achieve. It's not necessary the case that pursuing that utility function

        would deliver what is the goal of utilitarianism (maximize the happiness of everybody).

        - Can you imagine a system that acts according to an utilitarianist utility function? It's unworkable because a system should be capable to compute all possible consequences across time among possible people —> this would require a vast amount (too vast) of knowledge. (A system should have to be 'Super Intelligent').
        - Reference. Book of Nick Bostrom called 'Superintelligent'
    - Should AI systems adopt an utilitaristic reward function?

Issues with act utilitarianism

- Does it provide a good decision procedure
    - Can we choose what to do by optimising the outcome our actions? Do we have the information to make this calculation? Can an AI system have the information?
        - If we have information about the outcome of our actions we should take them into account, but what if we do not have exhaustive information about the outcome of our actions? Should we rely on probability? Can we use the consequences of an action as a standard way to process the actions?
- Does is provide a good standard for assessing decisions?
- What is the link between utility and a reward function?

Act utilitarianism: Problems

Do the best action that gives a tradeoff between the happiness and the pain on everybody.

- **Is it too demanding.**
    - Should I give to the poor all that I above the minimum that allows me to survive?
    - Should I give the same importance to everybody, regardless of their connection to me?
    - Reference: book called 'Machines like me' of Ian McEwan
- Is it OK to harm some people for the greater benefit of others

        - Reprisals? Torture?Sadism?
            - I would also think about e.g. racial laws: provided that "the utility of everybody should be taken into account equally", utilitarianism could easily hurt minorities, since actions "hurting" few people could be considered moral if they produce "happiness" for the majority.
- What could an utilitarian say:
    - The cases in which utilitarianism seems to fail are not realistic
    - There is no real contrast between utilitarianism and mainstream moral beliefs

Readings:

- Machine like me, Ian McEwan
- Nick Bostrom, super intelligence

Rule utilitarianism

'Utilitarian promise'

- an action is morally right just because it is required by an optimific social rule (a social rule the general compliance with which would provide the highest utility)
    - It is ok to tell the truth, not to steal, etc. since the general compliance with such norms would deliver the greatest utility
    - What about those exceptional cases in which the rule does not deliver What is you know that most people are not following the rule.
        - Should we be honest if most people around as are dishonest?

A further issue: distribution

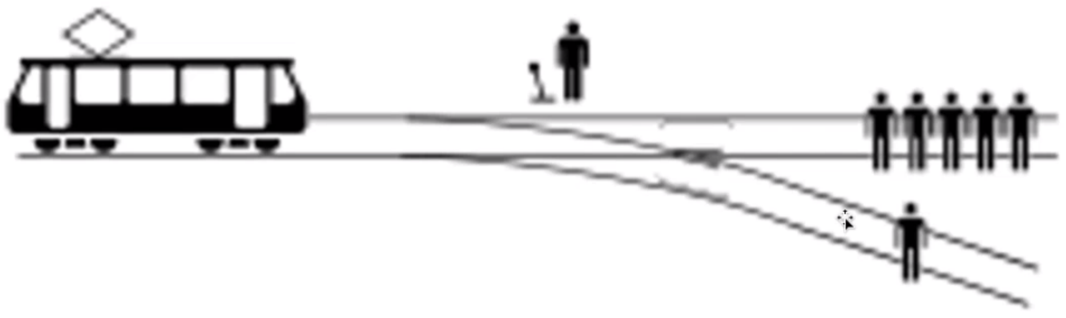
- **Does it matter how the good and bad outcomes are distributed**?
    - It is ok to make an action that benefits some to the detriment of others?
        - Should we also do actions that benefit everybody? Is possible to act to benefit everybody? We can't benefit everybody all the times. This is particularly true in political decisions.
    - Always if the benefits outweigh disadvantages?
- **Utilitarianism vs wealth maximisation**

  - Does utilitarianism coincides with maximizing the national product of a country? Are the choices that maximize the GDP coincident with the choices that maximize the happiness of the most people or not? Does it count only how the slice is big or also how the slices are created, from utilitarianism point of view? If you are an utilitarian, do you care only of the GDP or you also of the distribution?
    - From an utilitarian perspective, what counts is the total amount of happiness. So if increasing the GDP the slice of nobody is diminished, then this is better from the utilitarian perspective. If the slice of some are diminished we have to understand how this decreasing of happiness is compensated by increasing the happiness of others. In general utilitarianist favor the **redistribution of wealth, r**aising on the idea that is better a **great amount of money is used by a great amount of money, rather than if it's used only by one rich person.**
  - Utilitarianism favours (modest) redistribution of wealth, since the **same amount of money gives more utility to the poor than to the reach**
  - The impact of redistribution on wealth generation however has to be considered
- Wealth maximisation (adopted by some economic approach) aims at maximising the wealth in society regardless of distribution.
  - if we are utilitarian it matters how the slices of an apple are made
  - If by increasing the GDP the slice of nobody is decreasing this is good from utilitarianism. In general utilitarianism favors redistribution of wealth.

The trolley problem

Question: Imagine that you a train (a trolley) that is proceeding on trucks and if we do not do nothing the train will go straight and kill the 4 people that are on the trucks. You have the possibility to move a level and if you move the level the trolley is going to take the other direction and is going to kill one people.

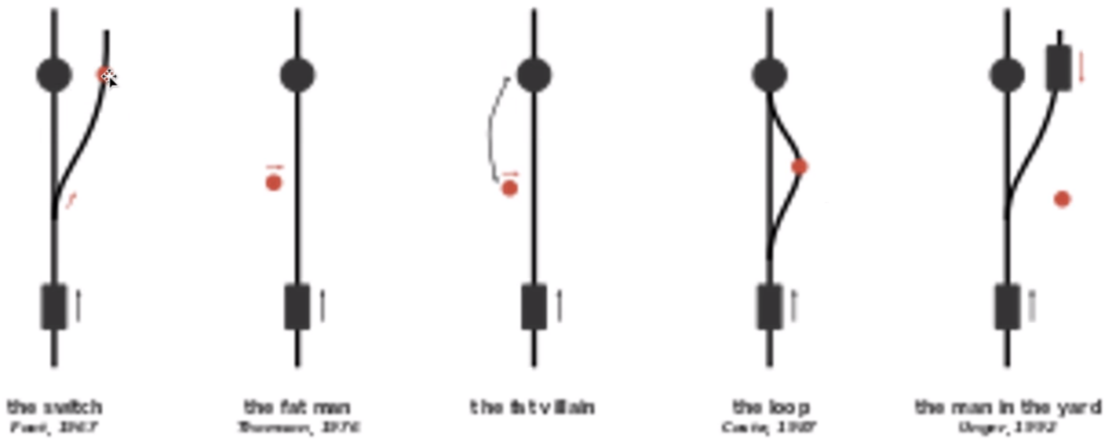
So, what is the right approach? We assume to adopt an utilitarian perspective



What would you do? What should an AI system tasked with monitoring traffic do

The utilitarian would push the level or let the trolley goes? **The utilitarian would push the level.**

- There is the obligation to not to kill so if in this case i kill one instead four i will not violate this moral obligation.
- Even the prohibition to kill is not an absolute duty. In this example we have to take an action that results in the death of a person, you knowing that this result is going to take place.
- The idea that we should never do anything that is against our duty is **deontologism. Deontologism** is the idea that morality consists in following our duties. If our duty is not to kill, i do not push the level (differently from the utilitarian perspective).



Black dot = 4 people.
Red dot = one additional person

- **1st case = the same as in the previous example**
    - In the 1st case how would you **justify** your action? None of us has chose to let the train kill the 4 people, but how you would justify the action?
        - We can say "Is better that 1 person dies rather than 4, I go for the better outcome"; this is the **utilitarian answer**.
        - Another answer could be "it's true, I pull the level but not with the intent to kill a person, but only to kill 1 instead 4." The kind of reasoning behind this answer refers to the **double effect principle:** (this is not an utilitarian aspect) when doing an action we have to distinguish the intended purpose of the action and the side effect of the action. So we may say: in this case my action has the purpose of saving 4, killing 1 was a side effect, so it was **not my intention**, therefore the action is ok. So, behind the double effect principle there is **the distinction between the intentional objective and the side effect.**
            - Utilitarians do not make this distinction: for them we have to take into account all the effects of our actions, regardless the fact that is intentional or not, you have only to reach the best outcome among all the possible.
- **2nd case**: **fat man case**. There is a fat man, you are near the fat man. The trolley is going to kill the four people, but you push the fat man on the trolley it would be so fat to stop the trolley. **Push the fat man on the tracks or not to do it, what would you do?**
    - **A utilitarian would push**.
    - Difference between the 2 cases (1st and 2nd): if you are an utilitarian no problem to push. But if you are not a utilitarian you may wonder because there is a difference in this case: you are intentionally performing harming someone, but this happens also when you push the lever. **Difference:** en ethicist would say that in the 2nd case you **use the fat person as a mean to achieve your goal of saving the four.** if you push the lever what you are trying to do is saving the people and the death of 1 man is not my intention. If you put the fat man you certainly have the intention to put the fat man there, so harming the fat man is your choice-objective because thanks to this objectives you don't kill the 4 people.
    - If you push the lever you may say that what you are trying to do is to save the 4 people by sending the train in another direction and the killing of the

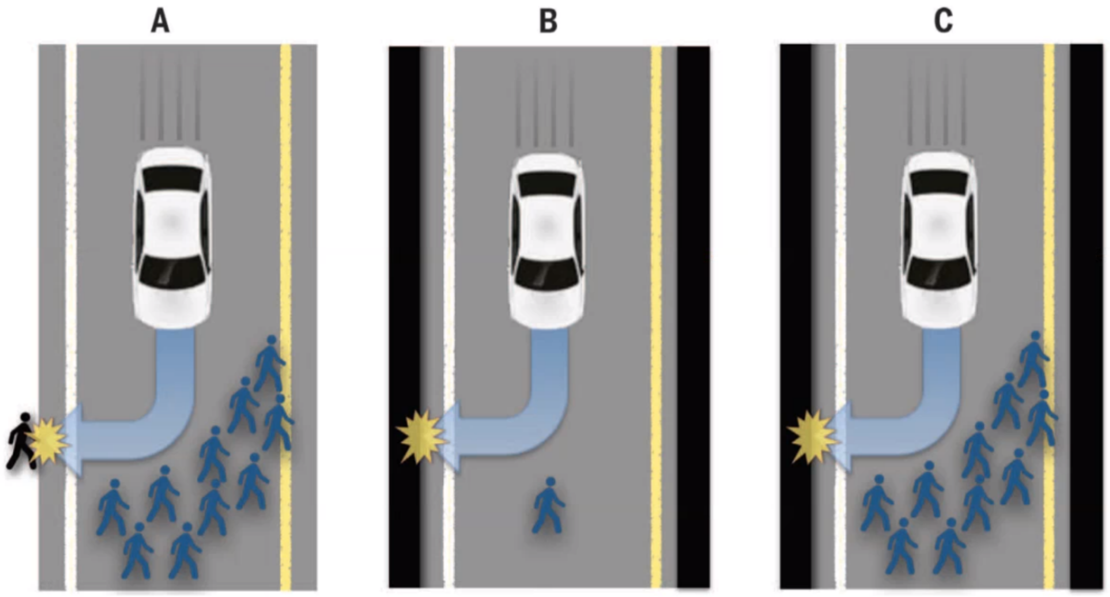
other alone man is a side effect that was not in our intentions. If you push the fat man then certainly you have the intention of pushing the fat man on the train line and making so the its body stops the train. So harming the fat man in this case is our choice -objective because thanks to this objective we result in our goal of saving the 4 people.

- **3d case: fat man is also bad man** (criminal which want to kill the 4 people) You have the option of pushing the criminal and saving the people that the criminal himself is destinated to death. Probably in this case you would have less problems in pushing the fat man.

These examples show that utilitarianism can be problematic cannot fit our common sense reasoning in most of the applications.

The social dilemma of autonomous vehicles

Question: What should an autonomous vehicle do in situations in which harm is unavoidable? There is no time to break, what should the vehicle do?



Bonnefon et al. 2016

Situation A

What would you do if you were a driver and what would you do if you were a programmer? Imagine what you could do begin the driver or the programmer that can program in advance what the car should do in this situation.

- **Drive**r: as a driver, turning right and killing a single people would be the same as pulling the lever (not pushing the poor fat man)-from a moral perspective. It's called **state of necessity**.
    - **From a utilitarian perspective would it be relevant that people are crossing with red light**? Not.
        - From the **moral perspective** we can say that they decide to violate the law so it's right to kill the people that are crossing the road, rather then those who are not violating the law
- **Programmer: i**nsert the instruction to go against the pedestrian because this choice minimizes the number of deaths. Do the choice that minimizes the number of deaths.

Situation B

If you turn you will likely die on the wall. There is just 1 person in the car.

From an utilitarian perspective what would you say?

- **Driver**: would you straight ahead or turn? The situation is between our life as a driver and the life of the pedestrian.
    - From an **utilitarian perspective.** Apply expected utility (50% io, 50% pedestrian). From an utilitarian perspective you would apply expected utility.
    - From our common sense we would prefer our safety. Many of us would say that in such cases is better make the preference on ourselves and kill the person.
- **Programmer:** —non si è capito—

In the **law** if you go straight you invoke the state of necessity (to avoid harm to yourself) so you not be accused for omicide.

Judith Jarvis Thomson: The surgeon case

- A brilliant transplant surgeon has five patients, each in need of a different organ, each of whom will die without that organ. Unfortunately, no organs are available to perform any of these five transplant operations.
- A healthy young traveler, just passing through the city in which the doctor works, comes in for a routine checkup. In the course of doing the checkup, the doctor discovers that his organs are compatible with all five of his dying patients.

- Suppose further that if the young man were to disappear, no one would suspect the doctor. Do you support the morality of the doctor to kill that tourist and provide his healthy organs to those five dying people and save their lives?

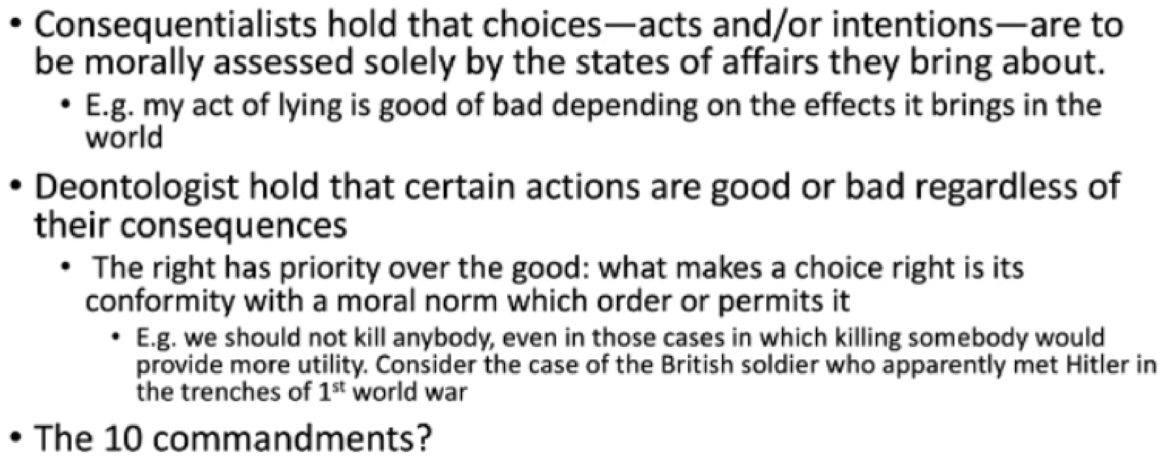
Supposing that in the future we have robots that make choices.

What would an utilitarian doctor would do?

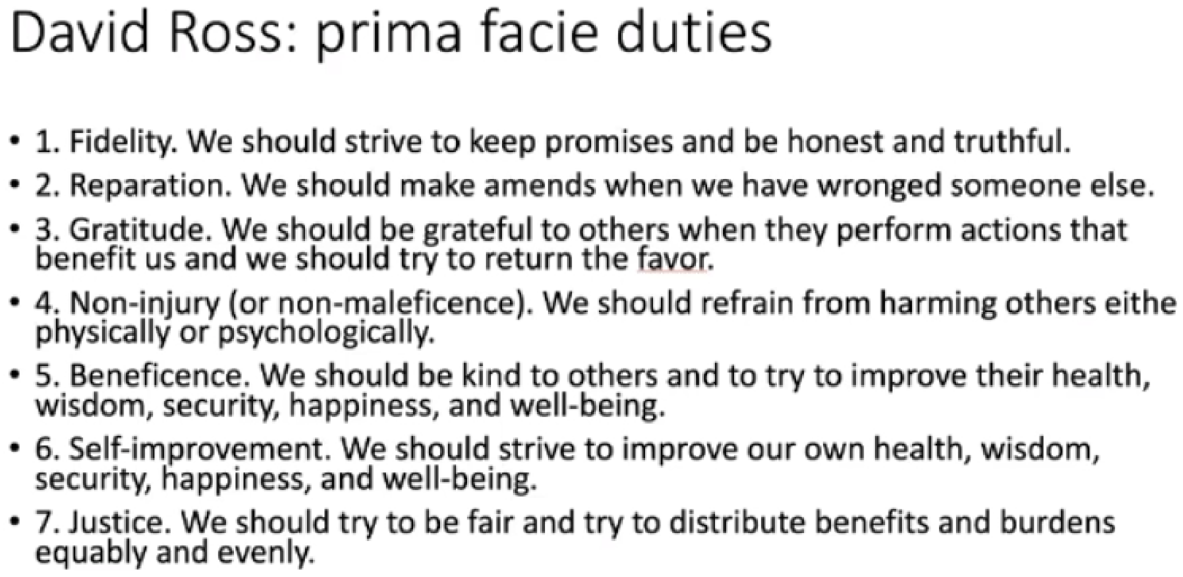
- An utilitarian would say that this example is not realistic, it is not so likely and possible rule-utilitarian would say that is better to sacrifice a young man to save 5 lives.

Deontology

This is an alternative to utilitarianism.

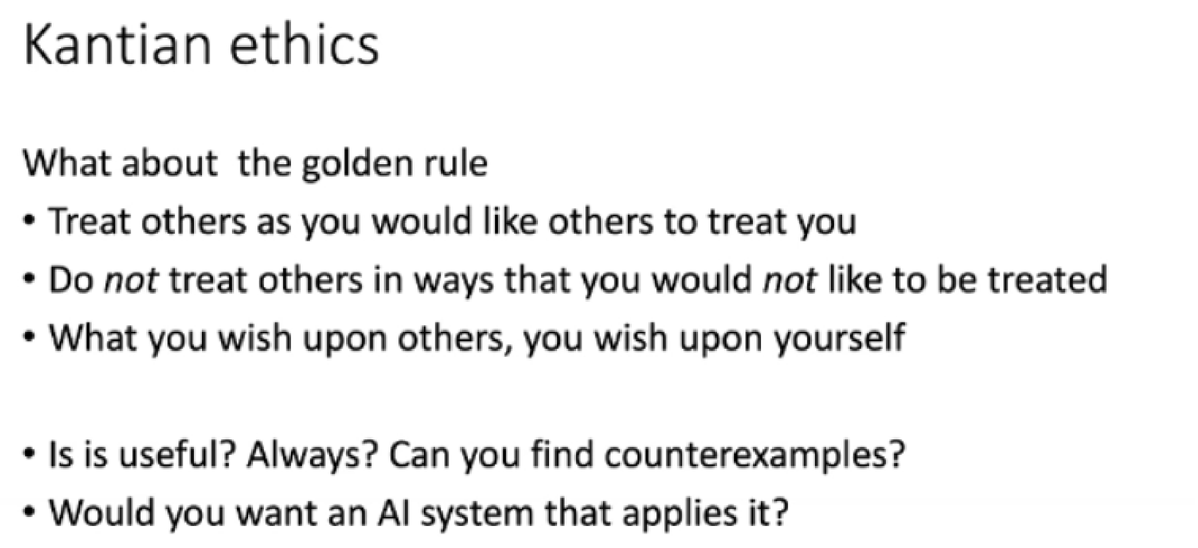


- Consequentialists hold that choices—acts and/or intentions—are to be morally assessed solely by the states of affairs they bring about.
  - E.g. my act of lying is good of bad depending on the effects it brings in the world
- Deontologist hold that certain actions are good or bad regardless of their consequences
  - The right has priority over the good: what makes a choice right is its conformity with a moral norm which order or permits it
    - E.g. we should not kill anybody, even in those cases in which killing somebody would provide more utility. Consider the case of the British soldier who apparently met Hitler in the trenches of 1st world war
- The 10 commandments?



David Ross: prima facie duties

- 1. Fidelity. We should strive to keep promises and be honest and truthful.
- 2. Reparation. We should make amends when we have wronged someone else.
- 3. Gratitude. We should be grateful to others when they perform actions that benefit us and we should try to return the favor.
- 4. Non-injury (or non-maleficence). We should refrain from harming others eithe physically or psychologically.
- 5. Beneficence. We should be kind to others and to try to improve their health, wisdom, security, happiness, and well-being.
- 6. Self-improvement. We should strive to improve our own health, wisdom, security, happiness, and well-being.
- 7. Justice. We should try to be fair and try to distribute benefits and burdens equably and evenly.

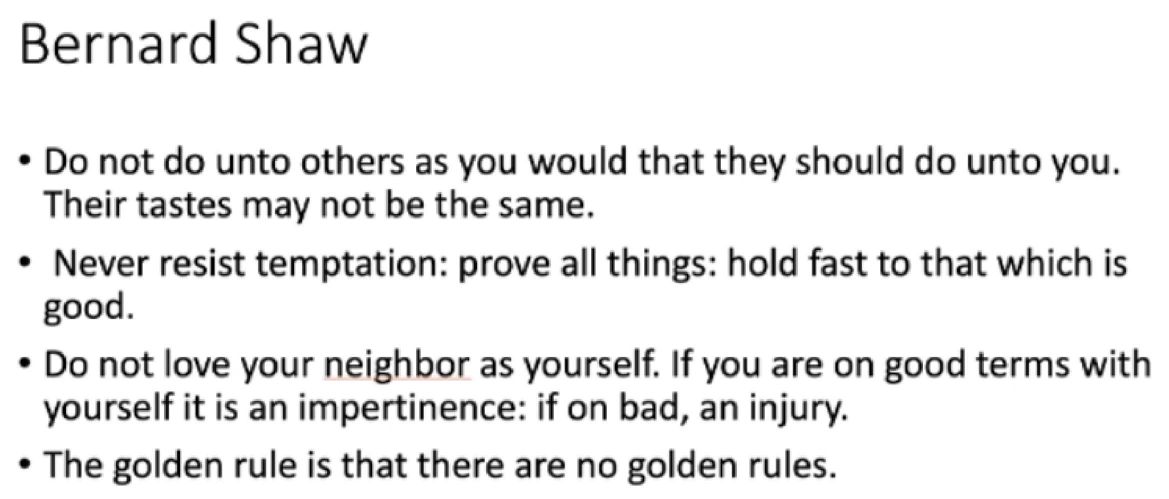


Kantian ethics

What about the golden rule

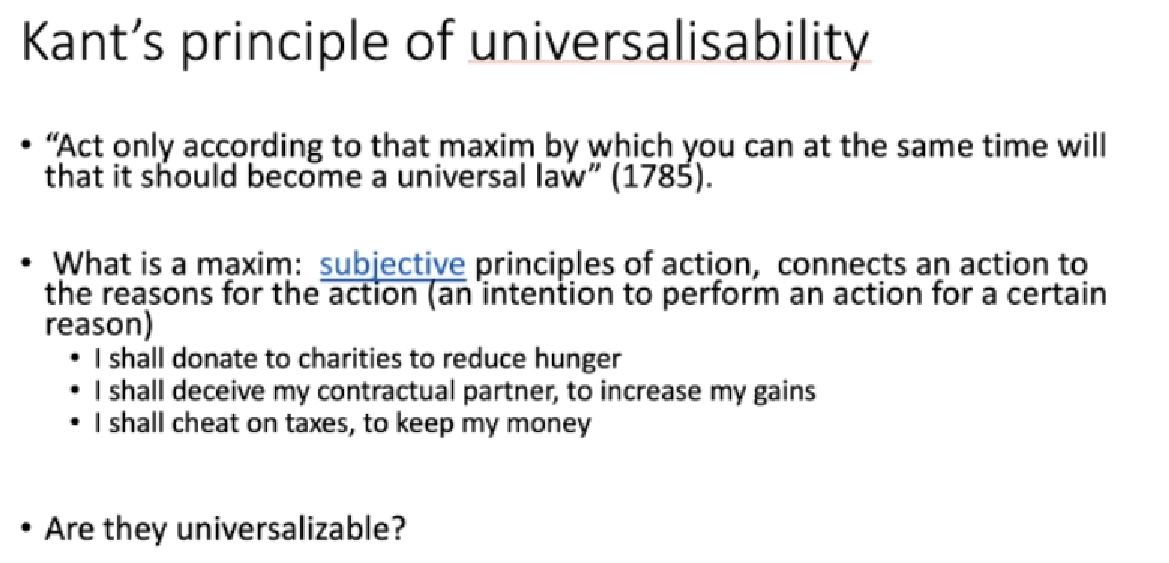
- Treat others as you would like others to treat you
- Do *not* treat others in ways that you would *not* like to be treated
- What you wish upon others, you wish upon yourself

- Is is useful? Always? Can you find counterexamples?
- Would you want an AI system that applies it?



Bernard Shaw

- Do not do unto others as you would that they should do unto you. Their tastes may not be the same.
- Never resist temptation: prove all things: hold fast to that which is good.
- Do not love your neighbor as yourself. If you are on good terms with yourself it is an impertinence: if on bad, an injury.
- The golden rule is that there are no golden rules.



Kant’s principle of universalisability

- “Act only according to that maxim by which you can at the same time will that it should become a universal law” (1785).

- What is a maxim: subjective principles of action, connects an action to the reasons for the action (an intention to perform an action for a certain reason)
  - I shall donate to charities to reduce hunger
  - I shall deceive my contractual partner, to increase my gains
  - I shall cheat on taxes, to keep my money

- Are they universalizable?

If I am going to put ethics inside an AI system, which ethical approach should i try to put in the AI system?