# Lesson 4

# A) Deontology

# 1) Kantian ethics

Last time we approached **consequentialist** approaches: to address if an action is correct we have to see which is its consequence.

In particular, we have seen **utilitarianism**.

We have considered this idea to be attractive but to be problematic when in order to do good, we need to cause harm.

Consequentialism is an approach linked to decision theory: you have to maximize the utility of the outcome of the actions. See how good or bad can result from different choices and compare the costs of the choices.

The other approach is **deontology**. The basic idea

# Deontology

- 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
- Deontologist hold that certain actions are good or bad regardless of their consequences
  - Lying is always bad, regardless of its effect.
- The right has priority over the good: what makes a choice right is its conformity with a moral norm which order or permits it, rather than its good of bad effect.
  - E.g. we should not kill anybody, even in those cases in which killing somebody would provide more utility. Is this always the case
     Consider the case of the British soldier who apparently met Hitler in the trenches of 1<sup>st</sup> world war
    - Consider the case of the British soldier who apparently met Hitler in the trenches of 1<sup>st</sup> we
      What would a rule utilitarian say in such a case?
- The 10 commandments?

# For deontologists, doing good actions has priority over the good, so over the outcome of the actions.

Soldiers should save Hitler or shoot him?

In Germany, after the case of twin towers, they told first that se c'era un aereo con civili che andava contro una struttura allora l'esercito era autorizzato a sparare

all'aereo. Successivamente si decise il contrario.

Alex: I have a question about judges at tribunals, their decisions regarding a case are "usually" deontological?

Sartor's reply: Judges have to comply with the laws of the states. However there are some situations, like the 'state of necessity'. We are considering ethical norms like 'no not harm people' 'do not kill anybody' etc

# Some ideas for being impartial

Ethics and impartiality

- Is ethics linked to ideas of fairness or impartiality?
- Is it unethical to have a preference for oneself (or one's friends)?

### 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 the golden rule useful

- Always? Can you find counterexamples?
- Would you want an AI system that applies it (with regard to its owner)?

What are the basic ideas of **deontology**?

 Religion → for example, the golden rule is a moral principle which denotes that you should treat others the way you want to be treated yourself. For example, the golden rule suggests that if you would like people to treat you with respect, then you should make sure to treat them with respect too.

Some ideas of deontological approach can be found also in the so called 'golden rules'.

Idea of impartiality is at the basis of deontology.

We focus on the situation of **individuals** here —> the golden rule is different from the one of utilitarianism.

Here the concept concerns actions.

Imagine that I am a masochist (be beaten by others) and I follow the golden rule i like to beaten them so i have to beat them even if they do no like it. So since people might be different preferences, the same rule could not work. Would you like a robot to behave according to the golden rule? Golden rule: if the robot is a representative of yours, imagine that if the others are honest with the robot we would have also robot honest with others, but if the they do not behave in this way, also the robot would not behave good. Is this correct?

# Whenever there is a mismatch between preferencies, the golden rule has some problems.

## **IMMANUEL KANT**

Let's focus on Immanuel Kant  $\rightarrow$  the prototype philosopher of deontology

Al should provide good to society and harm should be avoided —> Al tries to maximizes the social good —> consequentialism side in the European Document seen last time (Trustworthy Al).

But, regarding ethics in AI, there are also other ideas which have difference sources. One of them is the one of Kant.

# Immanuel Kant

- One of the greatest philosophers of all times
- Lived in Prussia (1724-1804)
- Addressed
  - The theory of knowledge: Critique of pure reason
  - The theory of morality: Critique of practical reasons
  - The theory of aesthetics (art): Critique of judgment
  - Law, logic, astronomy, etc.



# Kant's ethic and the principle of universalizability

- "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: a subjective principle of action, it 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
  - I shall tell the truth, to provide trust
- Are they universalizable? Would I want them to become universal laws, that are applied by everybody?

Kant considered that not all maxims that we can adopt are **universalizable**: what does it means?

## An universalisation test

- Shafer Landau. The test of universalizability:
  - Formulate your maxim clearly state what you intend to do, and why you intend to do it.
  - Imagine a world in which everyone supports and acts on your maxim.
  - Then ask: Can the goal of my action be achieved in such a world?
- The process ensure some kind of fairness

Apply this principle to

- Cheating in an exam, in order go get a good mark
- Giving money to a charity to relieve
- Would we want a robot following this maxim?

You have to formulate the maxim of your actions clearly and imagine what would happen if everybody happen on the maxim. To see whether the maxim is universalizable you have to imagine that everyone act following this maxim —> for example if everyone cheat during the exam, taking the exam becomes impossible.

Robot in e-commerce domain: a robot has to decide whether to cheat its partner (take the money and run away);by applying the universalization test the robot

consider a scenario where each agent cheats  $\rightarrow$  e-commerce collapse  $\rightarrow$  the maxim of cheating your partner is not universalizable

# Immanuel Kant vs Benjamin Constant

- Should one must (if asked) tell a known murderer the location of his prey.
  - It is ok to refuse to answer?
  - It is ok to tell a lie (e.g., if threatened by the murderer)?
- Is the maxim of telling lies universalizable?
- Is it defeasible?
- Its it Ok to have a robot that tells lies:
  - What about Asimov Liar
  - What about HAL in



The murder asks to say where there is the person who is hidden in my house.

For Kant is ok to refuse the answer? Yes, it is ok because the maxim 'do not say anything'

However, there are situation in which you cannot simply say nothing, but you are forced to answer something (e.g. l'assassino ti sta intimando di parlare)

Is it ok for a robot to tell lies or not? (Asimov "Liar!" , short story <a href="https://en.wikipedia.org/wiki/Liar!\_(short\_story)">https://en.wikipedia.org/wiki/Liar!\_(short\_story)</a>)

Should an AI system always tell humans what the system thinks to be truth or not necessarily? Let us assume that we can use AI system to distribute lies.

Lorello: Truth or falsity require consciousness. The only discriminant, with today's technology at least, would be "explainable answer"/"non explainable answer"

Biniam: I would argue that the maxim of saving someones life is greater than the maxim of lying. To use your previous example

#### What is rational?

For Kant rationality includes also the requirements of Universalizability  $\rightarrow$  Practical Rationality.

For example, it is not rational to cheat for achieving my goal, because it is not universable! This is different from **instrumental rationality**, where you have to

#### act in the way you maximize the probability of achieving a given goal.

Again, if we talk about spread fake news online and this make me reacher, this would be rationale in instrumental rationality, but according to Kant this would be rationale —> i would not be morally ethical because i would act according to a maxim that is not universalizable.

Is rationality related to morality?

There are two viewpoints:

A. morality is strictly connected to rationality: to be rational you have to be also a good person

B. You can be equally rationale if you are criminal or you are good person

## For Kant morality overcome any other needs, in the sense that you cannot be rational without acting morally.

A crucial point is Kant approach to morality is the opposition between **hypothetical imperatives** (instrumental rationality) **and categorical imperatives** 

# Hypothetical imperatives

- Hypothetical imperative: they require us to do what fits our goals
  - I would like to have more money
  - If cheat on taxes I will have more money
  - · I shall cheat on taxes to have more money
  - I would like to get a good mark
  - If I study I will get a good mark
  - I shall study
  - Is this OK?
  - The imperative is dependent on what I want (getting good marks, having more money)
    - I shall cheat on taxes, to having more money!

Kant observes that the imperative depends on what I want, and to find something that holds for everybody we have to go through **categorical imperative**.

# The categorical imperative

- A moral imperative that applies to all rational beings, irrespective of their personal wants and desires,
- "Act only on that maxim through which you can at the same time will that it should become a universal law"
  - · make false premises when it suits you to do so?
  - · refuse help to do those who are in need when it suits you to do so?

Look at the two proposals (pointed list), none of them satisfies the categorical imperative:

If everybody would act according to the first maxim the full institution of premises would collapse.

The second maxim is more debatable, but for Kant also this maxim would not pass the test provided by the categorical imperative, because if it were the case nobody would help you and you would not accept this kind of state of the affairs.

Act in a certain way only if my maxim can be followed by everybody.

I must act on the maxim (do not lie to increase trust in society for example) if we believe that while we acting like this, the maxim can become a categorical imperative.

# The good will

- The morality of an action only depends only to the extent that this action is motivate by our good will, i.e., by the necessity to comply with the categorical imperative
  - E.g., if I do well my job only in order to get a promotion, or be better paid I am not acting morally
  - I am acting morally if I do well my job because I think that this is my categorical duty, since I believe that everybody should act upon the maxim that they ought to do well their job to ensure societal progress
- The good will is the only thing that is good in itself
  - Do you agree?

For Kant we may act well in order to achieve certain goals: for instance I may be kind to a colleague or to a friend because i want that he/she would be kind to myself. Or I may be honest because I want to keep a good reputation. (example of good reviews in e-commerce)

For Kant my action is really good only to the extent which corresponds to a good intention —> the good will. I need to do the action only if i think that this is the action that everybody should do. For him the good will is the only thing that is good in itself.

The 2 examples in the slide:

- For Kant If I do my job well to be better paid this is not moral.
- In the second example I am acting morally to Kant.

# Another version of the categorical imperative: the principle of humanity

- So act that you treat humanity in your own person and in the person of everyone else always at the same time as an end and never merely as means
  - How is it linked to universalizability: As you consider your self as an end, you should consider the others in the same way (universalizability)?
- What does it mean treating somebody as an end (not as a mere means)
  - It cannot mean that we never use people for our purposes (e.g., when we ask for favours or pay for jobs)
  - It must mean that we should never treat people ONLY as means, without considering their values and purposes

How principle of humanity is linked to principle of universalizability?

- For Kant giving false information to people violates the principle of humanity. When cheating people I use them merely.
- If I am interacting with you for the purpose of getting some benefits (sell you an object in order to get from you some money) and I recognize that for you is better to get the object and for me is better to get from you the money and sell you the object. I'm not using you as an object, so it is okay.

# When does AI treat people only as means

- Autonomous weapons?
- Deceiving advertisements?
- Discriminatory appointments?
- When does AI fail to recognise humans as valuable entities, that should achieve their aims according to their choices?
- Can we treat AI systems only as means?

# Dignity

- For Kant rational beings, capable of morality (humans) have a special status "an intrinsic worth, i.e., dignity," which makes them valuable' "above all price
  - Because of dignity they deserve respect
  - They cannot be treated as mere ends
- What does it mean that AI systems should respect human dignity, respect humans

Humans, because of their dignity deserve respect. What does it means that AI should respect humans and human dignity?

Let us consider surveillance people that track people using face recognition even in the open spaces, would such a system respect human dignity? We may have different answers.

Another example: having systems that perform sentiment analysis or that detect people's behaviour and attitudes of people by looking at theirs faces. An AI system that is able to extract from people behaviour thoughts or attitudes may be not respectful of human dignity. —> This can be a way of disregarding human dignity. This is an important issue that comes up in Kantian approach.

According to Kantian reasoning humans deserve dignity because:

# The foundations of dignity

- Why do humans deserve dignity. Because they have
  - Reason: they act on reasons and are aware of this
  - Autonomy: the can choose what to do, and in particular to follow the categorical imperative rather than their subjective preference
- The kingdom of ends
  - In the kingdom of ends everything has either a price or a dignity. Whatever has a price can be replaced by something else as its equivalent; on the other hand, whatever is above all price, and therefore admits of no equivalent, has a dignity
- · What if AI system also had reason and autonomy
- · Would they become citizens of the kingdom of ends

What is dignity? What is the kind of respect that humans deserve? How can Al system affect human dignity?

**Personalized advertising** is consistent with the idea of dignity and autonomy? What if advertising is aggressive or misleading? For example, an advertisement that pushes me on gaming addiction on the basis of its knowledge about my weaknesses, is it ok?

# Morality as an aspect of rationality

- For Kant if we follow rationality, we have to be moral.
  - Can there be a rational criminal?
  - It is rational to pursue my wellbeing at the expense of others?
  - Is it rational for a company to develop a system that is profitable, but that will cause more harm than good (e.g.,

A criminal can be rational but we are again in the discussion of different form of rationality:

rationality is the ability to pursue a certain goal vs rationality as the ability of acting according to a categorical imperative

 $\rightarrow$  Kant would say that criminal would be bad and not fully rational.

Is an altruist irrational? Non lo ha detto

# Rationality and consistency

- 1. If you are rational, then you are consistent.
- 2. If you are consistent, then you obey the principle of universalizability.
- 3. If you obey the principle of universalizability, then you act morally.
- 4. Therefore, if you are rational, then you act morally.
- 5. Therefore, if you act immorally, then you are irrational.

What kind of consistency is this?

- If I deserve something no less than others, and I want it for me, I should recognise it also to others!
- Is this consistent with rationality? Is it required by it? Can I be rational, and pursue my goal to the detriment of other

According to Kant:

practical rationality  $\rightarrow$  acting according to categorical imperative  $\rightarrow$  acting according to maxims that you think that would be universal law = followed by everybody.

Being consistent = not make preference for yourself, but act according to a maxim that you think could be followed by everybody

### Issues of Kantian reasoning:

## Issues

- Does the principle of universalizability always provide acceptable outcomes
- Is it sufficient that the maxim of my action is such that I would like it to be universalised for this maxim to be good?
- Can you think of some examples when this is not the case?
  - Lying ? Robbing? Celibacy? Genocide?

We may wonder whether the principle of rationality always provides acceptable outcomes.

### **Principle of Generic Consistency**

Alan Gewirth tried to develop a system of morality based on Kantian ethics. The basic idea is that if I want something for me and I believe that I have the right of pursuing my goal, since I am not difference from others, also others have this right.

# Alan Gewirth (1912-2004): principle of generic consistency

- 1. I do (or intend to do) X voluntarily for a purpose E that I have chosen.
- 2. E is good
- 3. There are generic needs of agency.
- 4. My having the generic needs is good *for* my achieving E *whatever* E *might be*  $\equiv$  My having the generic needs is categorically instrumentally good for me.<sup>13</sup>
- 5. I categorically instrumentally ought to pursue my having the generic needs.
- 6. Other agents categorically ought not to interfere with my having the generic needs *against my will*, and ought to aid me to secure the generic needs when I cannot do so by my own unaided efforts *if I so wish*,
- 7. I am an agent  $\rightarrow$  I have the generic rights.
- 8. All agents have the generic rights.

Other attempts exist to develop a Kantian ethics.

La spiegazione di sartor è peggio di qualsiasi altra cosa mai udita. Quindi ecco qui la pagina di wiki (<u>https://en.wikipedia.org/wiki/Alan\_Gewirth</u>), ma non è comunque la cosa più digeribile di questa terra.

There are other approaches

# Approaches to universalisability

- Richard Hare (1919-2002)
  - Moral judgment are universalizable: the judgment that an action is morally right/wrong commits me to accept that all relevantly similar action are wrong
  - Moral judgments are universalizable in the sense that they take into account the satisfaction of everybody's preferences (back to utilitarianism)

### Christine Korsgaard (1952)

- My humanity (capacity to reflectively act from reasons) it to me a a source of value, and
- I must regard the humanity of others in the same way.

### **Richard Hare:**

Other approaches to universalisability, Hare tries to reconcile Kantian and Utilitarianistic approach. If i say that is morally right for me to cheat it commits that not only for me it is right to cheat but also for everybody to cheat. So, I can see that is not correct that everybody cheat and as I cannot accept that everybody cheat, than I cannot say that is right for me to cheat.

### **Recalling Utilitarianism:**

- Act Utilitarianism : action is moral if action has better consequences than all alternative actions I would take
- Rule Utilitarianism: action is right if it complies with a rule which compliance with everybody provide the best outcome.

Both ACT and RULE utilitarians must ASSUME NOTHING. They must actually poll or measure what act will produce the greatest utility.

The difference is that the ACT UTILITARIAN measures the consequences of a SINGLE ACT.

The RULE UTILITARIAN measures the consequences of the act repeated over and over again through time as if it were to be followed as a RULE whenever similar circumstances arise.

NOTHING is right or wrong in itself for a utilitarian. NOTHING! It all depends on the consequences of the act, the results are what matters not

### the act.

Fonte: <u>https://www.qcc.cuny.edu/socialsciences/ppecorino/intro\_text/Chapter 8</u> <u>Ethics/Utilitarianism.htm</u>

# Do we want Kantian robots

- Yes
  - They will be consistent
  - They will be impartial
- No
  - They may act on bad maxims
  - Their maxims may be too rigid

Having ideas of humanity, categorical imperative, impartiality, idea of dignity and autonomy are very important notions that come up when we talk about Ethics in AI.

## 2) David Ross

In moral philosophy, in addition to a short critique of Kantian ethics, his great accomplishment was the formulation of a major new ethical theory, a unique and still controversial system that combines deontological pluralism, ethical intuitionism, nonnaturalism, and so-called prima facie duties.

Ross's ethical system is deontological and anti-consequentialist since it is based on adherence to rules or duties rather than outcomes. It is pluralist in the sense that, unlike Kantian ethics and utilitarianism (monist systems based on a single, pre-eminent, all-encompassing rule or principle – namely the categorical imperative and the principle of utility, respectively), Ross recognizes several different fundamental rules or principles that he terms prima facie duties. Moreover – and this is a key element and a distinctive feature of his theory – he acknowledges that these duties can, and invariably do, collide and come into conflict with one another.

#### (https://iep.utm.edu/ross-wd/)

Here are the seven prima facies duties identified by Ross:

- 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 either 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.

# Defeasibility of duties

- Does it make sense to view duties as being defeasible?
- · Can we apply defeasible reasoning to reason with duties?
- Should an AI system admit exceptions to duties, or should it always ask humans?

# 3) Friedrich Nietzsche

Nietzsche had an extreme "anti-Kantian" approach, he believed that:

- The superior human (Übermensch) is beyond the traditional views of good and bad, beyond the morality of the herd
- One has duties only toward one's equals; toward beings of a lower rank, one may act as one sees fit, 'as one's heart dictates'
- The superior human does not find or discover values, he (or she) determines the values
- No need to be ratified; the only criterion of wrongness is 'that which is harmful to me is harmful as such'

Not so ok for AI systems (MADDAAAI -.-')

## Contractarianism

# Social contract theories

- In political theory:
  - A societal arrangement is just if it had (or would have had been) accepted by free and rational people
- In moral theory
  - actions are morally right just because they are permitted by rules that free, equal, and rational people would agree to live by, on the condition that others obey these rules as well (Shafer Landau)

In contractarianism we build morality on agreement: a moral rule is right if everybody have agreed to it.

## Hobbes

Hobbes argues that the state of nature is a miserable state of war in which none of our important human ends are reliably realizable. Happily, human nature also provides resources to escape this miserable condition, by developing a new state in which aggressions are punished by law.

#### Prisoner Dilemma in game theory

Prisoner B Prisoner A	Prisoner B stays silent (cooperates)	Prisoner B betrays ( <i>defects</i> )
Prisoner A stays silent ( <i>cooperates</i> )	Each serves 1 year	Prisoner A: 3 years Prisoner B: goes free
Prisoner A betrays ( <i>defects</i> )	Prisoner A: goes free Prisoner B: 3 years	Each serves 2 years

It is assumed that both prisoners understand the nature of the game, have no loyalty to each other, and will have no opportunity for retribution or reward outside the game. Regardless of what the other decides, each prisoner gets a higher reward by betraying the other ("defecting"). The reasoning involves an argument by <u>dilemma</u>: B will either cooperate or defect. If B cooperates, A should defect, because going free is better than serving 1 year. If B defects, A should also defect, because serving 2 years is better than serving 3. So either way, A should defect. Parallel reasoning will show that B should defect.

Because defection always results in a better payoff than cooperation regardless of the other player's choice, it is a <u>dominant strategy</u>. Mutual defection is the only strong <u>Nash equilibrium</u>

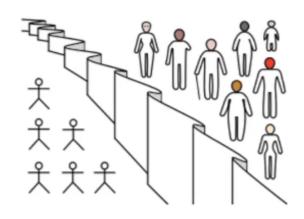
in the game (i.e. the only outcome from which each player could only do worse by unilaterally changing strategy). The dilemma, then, is that mutual cooperation yields a better outcome than mutual defection but is not the rational outcome because the choice to cooperate, from a self-interested perspective, is irrational.

## John Rawls (Book: A theory of justice)

In order to decide which rules are really impartial and moral, each individual has to reason under a **veil of ignorance,** i.e. without knowing their gender, social position etc. Otherwise, one may choose a rule according to their own interests (e.g. a rich person may want low taxes, while a poor one may want higher taxes to share the richness).

# John Rawls (1921-2002)

- · A theory of justice
- How to ensue that the social contract is fair?
- People should choose under a veil of ignorance, without knowing their gender, social position, interests talents, wealth, race, etc.



Rawls contends that the most rational choice for the parties in the original position are **two principles of justice**:

- First Principle (having priority): Each person has the same indefeasible claim to a fully adequate scheme of equal basic liberties, which scheme is compatible with the same scheme of liberties for all (liberty of conscience and freedom of association, freedom of speech and liberty of the person, right to vote, etc.;
- Second Principle: Social and economic inequalities are to satisfy two conditions:
  - They are to be attached to offices and positions open to all under conditions of *fair equality of opportunity*;
  - They are to be to the greatest benefit of the least-advantaged members of society (the difference principle). (JF, 42–43)

Would Rawls' approach be consistent with a society in which freedom of speech and association is strongly limited also by the usage of AI?

Nope, because it goes against the First Principle, which is more important of the second one.

Do the USA satisfy Rawls' idea of social justice?

Nope, because the difference principle is not satisfied! This happens also in other western countries, and the IT development has increased this issues by concentrating wealth on a limited number of individuals.

According to Rawls we have no rights to have a special treatment because of our talent: **Rawls' approach is anti-meritocratic.** Rawls claims that we have no moral merit for having been with certain talent; moreover, our talent is useful only because we leave in a society that values it (pillola concettuale tradotta: se fossimo nel neolitico, noi cervelloni saremmo delle pippe). However, for Rawls it is ok if society pays more for certain positions or give incentives to talented people as long as the difference principle is satisfied.

# Al in a just society (according to Rawls)

- · Does the deployment of AI in todays society fit Rawls' requirements
- · When may it conflict with the basic liberties?
- When with fair equality of opportunity?
- When with the difference principle?

## Juergan Habermas

He developed the Discourse Ethics, which refers to a type of argument that attempts to establish normative or ethical truths by examining the presuppositions of discourse. Variations of this argument have been used in the establishment of egalitarian ethics, as well as libertarian ethics. (<u>https://en.wikipedia.org/wiki/Discourse\_ethics</u>)

- A rule of action or choice is justified, and thus valid, only if all those affected by the rule or choice could accept it in a reasonable discourse.
- A norm is valid when the foreseeable consequences and side effects of its general observance for the interests and value orientations of each individual could be jointly accepted by all concerned without coercion
- The valid norms are those that would be the accepted outcome of an "ideal speech situation", in which all participants would be motivated solely by the desire to obtain a rational consensus and would evaluate each other's assertions solely on the basis of reason and evidence, being free of any physical and psychological coercion
- This approach assumes that people are able to engage in discourse and converge on the recognition of reasons for norms and choices

## Habermas and AI

- Would would we all agree if we engaged in an impartial discussion on how to use AI?
- Can we think of an AI system that engages in an impartial moral debate? What would it argue for?

# **B) Virtue Ethics**

- Ethics should not focus on norms nor on consequences
  - An act is morally right just because it is one that a virtuous person, acting in character, would do in that situation.
- · Ethics is a complex matter
  - Since there are many virtues, the right act is that that would result from the mix of the relevant virtues: honesty; loyalty; courage; impartiality, wisdom, fidelity, generosity, compassion, etc.
- Ethics cannot be learned though a set of rules, it application requires practical wisdom

Is it okay to teach morality to the AI through examples? Or should we use a rule based system

## Issues

- · How do we know what is virtues and what is not?
- How can we extract precise indications from an account of virtues and from virtuous examples? How much can we rely in tradition?
- What if virtues are in conflict?
- What are the paradigms of virtues to which we may refer to?

# AI and virtue ethics

- Should we, as developer of AI systems, be virtuous? What character traits should we cultivate in us?
- Should AI applications (AI agents be virtuous)?
- How can virtues be learned?
- If from example, I can the training of an AI system lead to a virtuous behaviour of it?

## Readings

- Shafer-Landau, R. (2018). The Fundamentals of Ethics. Oxford University Press.
- Singer, P. (2021). Ethics. In Encyclopedia Britannica: https://www.britannica.com/topic/ethics-philosophy

THE END.