The Ethics of AI

A PHILOSOPHICAL APPROACH TO ETHICAL ISSUES OF AI AND ROBOTICS

What is Ethics and what it deals with

- Ethics is the branch of knowledge that deals with moral principles. It is usually described as a branch of Philosophy.

- Ethicists want to:

1. Define a theory of the correct behavior.
2. Distinguish between right and wrong.
3. It deals with the practical life of humans: we must be able to judge if a specific behavior is right or wrong.

Branches of Ethics

Aristotelic: it holds that virtues are a behavioural model that benefits both the person and the society

Kantian: the only way to act ethically is following the concept of "duty". There must be no personal interest in the reasons of an action.

Utilitarianism: the goal of ethics is to provide the greatest happiness to the greatest number of people.

What problems does ethics deal with?

Ethics deals with questions in their "pure form", namely when all the other issues concerning the question are solved.

We have an ethical issue when "everything else works" and still there is something unacceptable.

Not an ethical problem

- A killer robot, because of a bug, makes a mistake and kills ten people.
- Solution? Fix that bug.
- So, we have to consider "ethical" those problems arising when things goes properly as expected and still something is wrong.

An ethical problem

- Videogames have become so immersive that many teenagers spend hours every day playing.

- In this case “everything else works”. In fact videogames are designed to be entertaining as possible. So we have an ethical issue.

Why ethics of AI

Every new innovation brings an alteration of the human political, social, relational, etc, context.

These modifications will change our concept of right and wrong, because they enable new possibilities.

Example: children surveillance.

Three fundamental fields of Ethics of AI

- Political issues
- Ethical dilemmas
- Social, relational and psychological issues

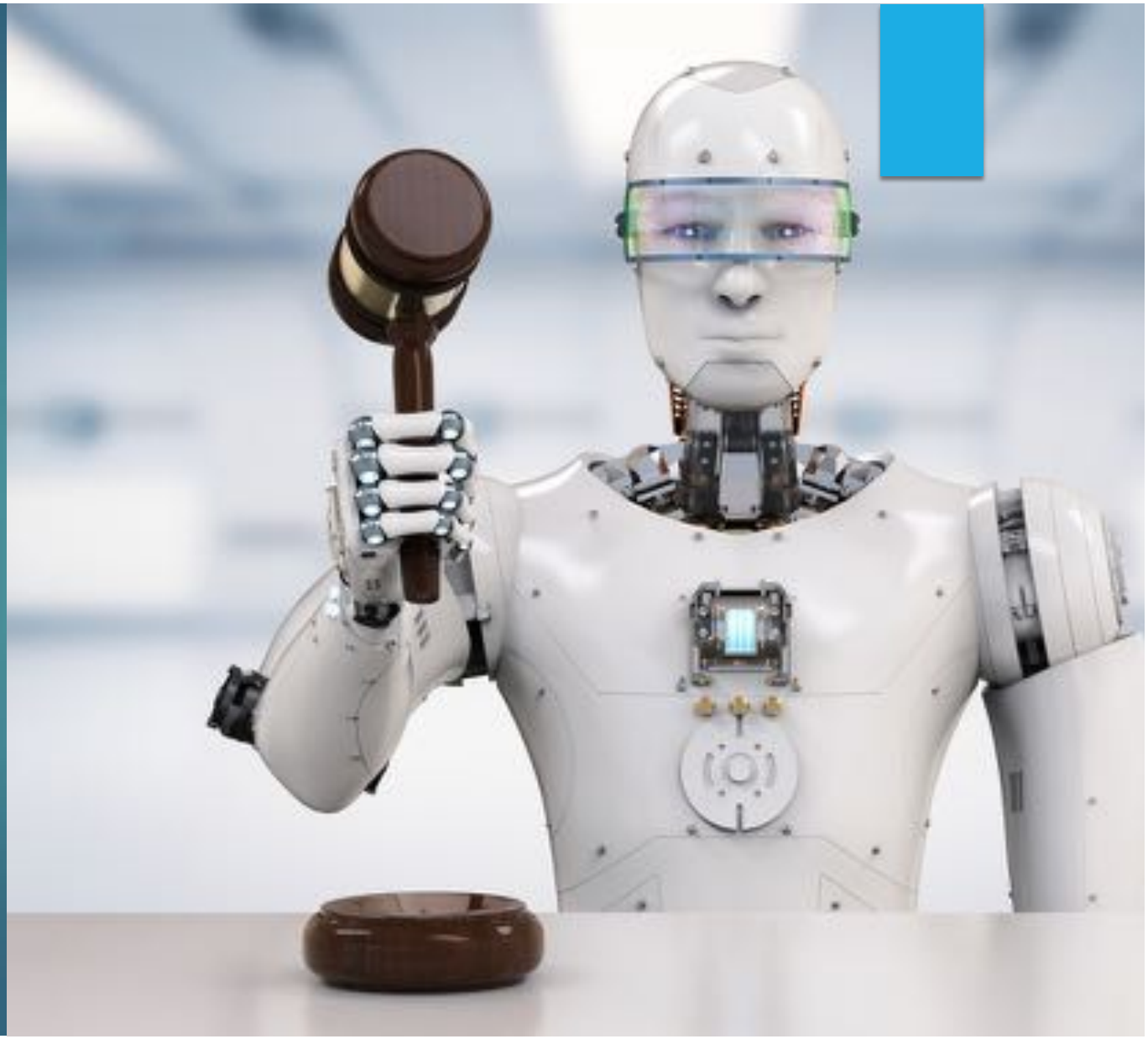
Unemployment

- Some studies (e.g. Frey & Osborne 2017) argue that automation will cause 40% of job loss in next 20 years.
- How to handle this problem? In future, will it be possible the automatization of nearly all works?
- How will our economy change to adapt to new technologies?

Frey, C. B., & Osborne, M. A. (2017). The future of employment: how susceptible are jobs to computerization?. *Technological Forecasting and Social Change*, *114*, 254-280.

A world without work?

- Will this increase the cases of depression?
- Then, what would be the role of humans?



AI Bias

- AI are not immune to human bias (Caliskan et al, 2017)
- "The reason why in recent times we are making steps forward so quickly is because we have become very good at mimicking human intelligence and put it in the artificial intelligences" (Joanna Bryson)

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186

Examples

- Tay: the nazi bot

- Family is for women, work is for men.

Wachter-Boettcher, S. (2017). *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*. WW Norton & Company.

Robots in warfare: How will wars change with robot soldiers?

- Robots are the perfect soldiers. (Arkin 2009)
- No fear
- No revenge
- No pressure
- No sadism

Arkin, R. C. (2009). Ethical robots in warfare. *IEEE Technology and Society Magazine*, *28*(1), 30-33

Ethical issues

Should it follow an unethical command?

The lieutenant's dilemma.

What is an ethical dilemma?

"An argument forcing a person to choose either of two ethically unacceptable alternatives."

No other alternatives (to do/not to do)

Each of the "horns" are unacceptable (To kill A/to kill B)

Characteristics

- There are good reasons (or bad reasons) for each horn.
- Anyway the decision will have a moral cost.

Unfortunately, in reality we cannot save them all.

Real dilemmas - False dilemmas

- **Real dilemma**
- There is a runaway trolley barreling down the railway tracks. Ahead, on the tracks, there are five people tied up and unable to move. The trolley is headed straight for them. You are standing some distance off in the train yard, next to a lever. If you pull this lever, the trolley will switch to a different set of tracks. However, you notice that there is one person tied up on the side track.
- You have two options:

- Do nothing, and the trolley kills the five people on the main track.

- Pull the lever, diverting the trolley onto the side track where it will kill one person.

- **False dilemma**
- Generally relies on one or more false assumptions.
- In Rome there is too much noise in night. You can regulate the night-life or let things as they are. If you regulate the night-life, city's life will change drastically in a negative way. If you do nothing, people cannot sleep.

- Assumption: the noise cannot be reduced without cut out the night life.

Ethical dilemmas are badly formulated questions

- If we have a real ethical dilemma we can do two things to "get out of the horns":

1. Accept that something unethical will be done.
2. Check again the assumptions of the argument.

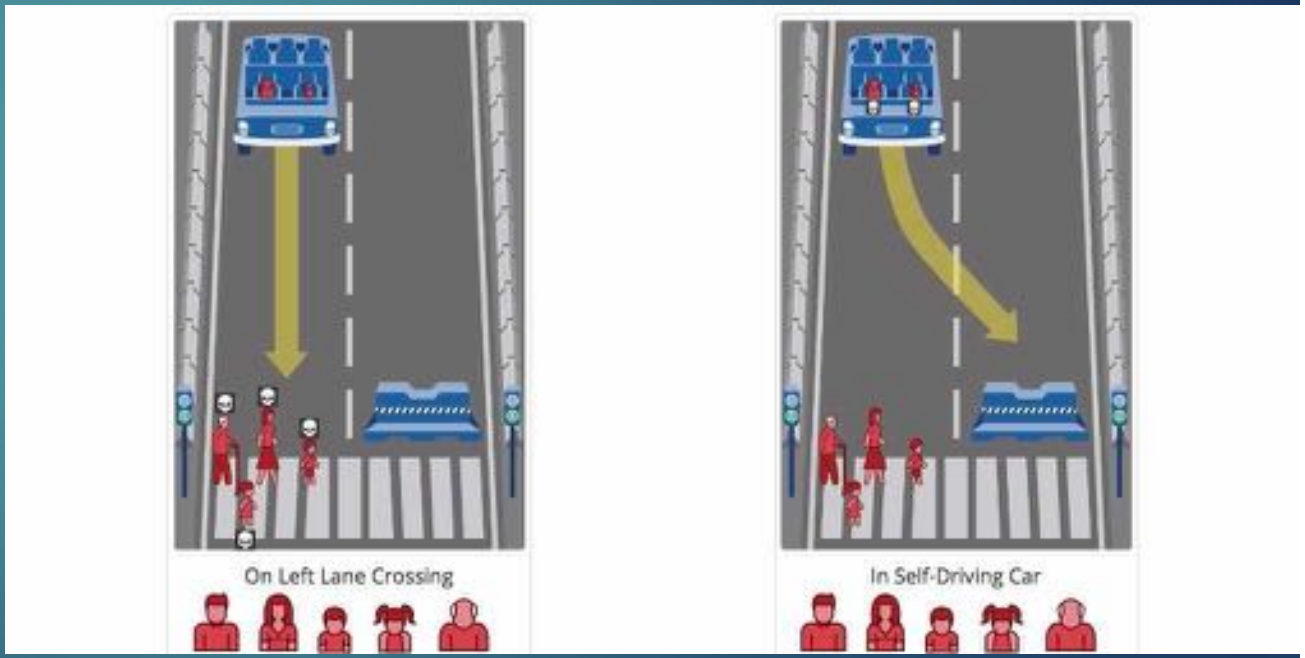
- Philosophy always tries to check the assumptions: dilemmas are seen as "badly formulated questions".

Autonomous driving cars ethical dilemma

- Why is different compared to humans?

Humans take a singular choice, here we design the universally valid choice.

- Utilitarian solution
- Life-value solution (social value)
- Evolutionary solution



How do we set up questionnaires?

- MIT gave questionnaires to people about this dilemma . They had to decide what to choose.

- Neuroscience demonstrated that if we make a decision:

- Rationally (if we have time): mainly utilitarian solutions
- Emotionally: various solutions. Neither altruistic nor egoistic

- So, the way we set up these enquires is fundamental.

Emdedded human behaviour

- What if we really behave like Spiderman?
- 1) Humans do not emotionally choose any horn of a dilemma.
- 2) Humans do not have the ability to understand, with certainty, that no other options are available.

Rebuild the question

Humans

- Not certain that it is a dilemma
- Psychological resistance
- Rational and emotional thinking mixed
- bad senses and bad comprehension of the environment

Autonomous driving cars (ADC)

- Certainty of dilemma
- No psychology
- No emotions
- Extremely higher calculation power of the environment

Out of the horns

1. ADC understands it is facing a dilemma.
2. It reduces randomly its calculation power of the environment.
3. It will stop when an acceptable ethical solution shows up.
4. We have avoided the ethical problem

Social, relational and psychological issues

- How our society, our psychology and our relation context could change because of the constant interaction with AI's and Robots with human-like behaviors?

Companion Robots

- Companion Robots (CRs) have the specific goal of building a relation with the human user.
- The final goal is to improve the user's quality of life.
- CRs are manly caring robots for elders or for children.

Main ethical concerns for CRs

Deception objection

The potential reduction of human contact

Hallucinatory danger for users

What if our relations are mainly, or only, with a robot?

The «projection» is a psychological concept involving the inconscious transfer of an emotion or a sentiment to something not directly related with the original object.

This is a normal process in all the relations, but with robot it could spread a lot.

Is it my wife or a robot? Turkle's elder

- A CR is assigned to an elder in a retirement home.
- In couple of few weeks he treated the CR as it was his ex-wife.
- The robot has surely accomplished the goal of user-friendliness but is this acceptable?

Turkle, S., Taggart, W., Kidd, C. D., and Dasté, O. 2006. Relational artifacts with children and elders: The complexities of cybercompanionship. Connection Science 18(4): 347-361.

Human relations

CRs doesn't have implicit relational restraints

These are a mix of cultural, experiential and psychological aspects that set what we consider acceptable or inacceptable in a relation.

In the case of Turkle's elder the efficiency of the CR in user-friendliness, becomes a loss in effectiveness in improving the quality of life.

Human constraints as a design principle

- Some of these issues are approachable considering human constraints as a design principle.
- Human sociality and relationality relies on the limits of human condition.
- To think ethically we should be aware that efficiency is not always effective.