TDT4136 Introduction to Artificial Intelligence

Lecture 10 - Part 2 Artificial Intelligence and Ethics

Pinar Özturk

Norwegian University of Science and Technology

2021

The Emergence of AI-Ethical Principles

In the last few years, a number of institutions have published AI principles:

- The Asilomar AI principles (Future of Life Institute, 2017)
- Principles for Algorithmic Transparency and Accountability (ACM 2017).
- IEEEs General Principles of Ethical Autonomous and Intelligent Systems (IEEE 2017)
- Five principles for a cross-sector AI code (UK House of Lords, 2018)
- AI ethics principles (Google, 2018)
- Ethics guidelines for trustworthy AI (European Commission, 2019)
- . . .

An older one: Asimov's Three Laws of Robotics

- A robot may not injure a human being or, through inaction, allow a human being to come to harm;
- A robot must obey the orders given it by human beings except where such orders would conflict with the First Law;
- A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

'Runaround' (1942), Isaac Asimov

Example: The 7 EU principles

- Human agency and oversight: AI systems should empower human beings, allowing them to make informed decisions
- Technical Robustness and safety: AI systems need to be resilient and secure. They need to be safe, ensuring a fall back plan in case something goes wrong
- Privacy and data governance: besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured
- Transparency: the data, system and AI business models should be transparent
- Diversity, non-discrimination and fairness: Unfair bias must be avoided
- Societal and environmental well-being: AI systems should benefit all human beings
- Accountability: Mechanisms should be put in place to ensure responsibility and accountability for AI systems

Common Grounds

There are many different lists of principles, but it seems that they all can be synthesized into five key principles

- autonomy (people should be able to make their own decisions, e.g. human-in-the-loop, privacy protection)
- beneficence (society at large should benefit)
- non-maleficence (harmful consequences should be avoided, e.g. systems should be robust)
- justice (diversity, non-discrimination and fairness)
- explicability (transperancy and explainability)

Problem with Principles

It is good to state principles! However they also create problems since they are very high-level.

- They can be interpreted in different ways.
  - For example, autonomous killer drones can be considered as being bene ficient for the soldiers, or being morally impermissible, because machines decide about life and death.
- .They can conflict with each other in concrete cases.
  - For example, privacy and data collection for health science can confict.
- They can come into confict in practice.
  - For example, an excellent diagnosis might still be preferable even if its. reasoning cannot be explained.

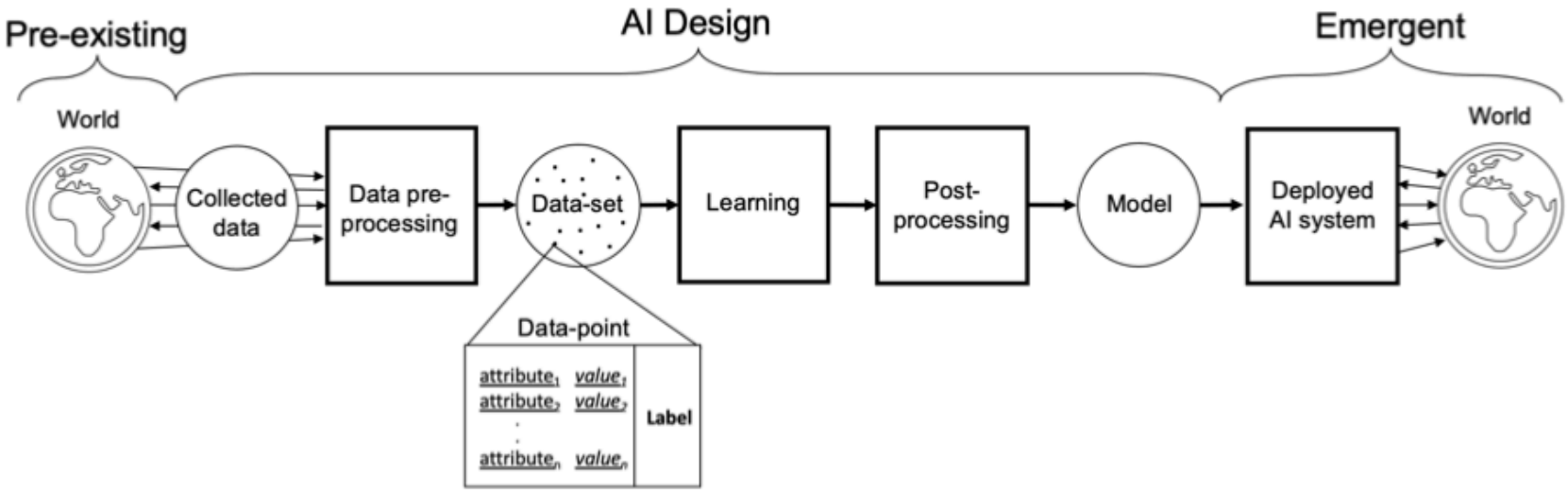
It is nevertheless good to have such principles as orientation points along one can evaluate solutions.

Artificial Intelligence $\neq$ Machine Learning

- In this course we looked at main stream AI, and not Machine learning
- Unbelievable many times ML is considered as if it is AI, but it is only a subfield of it
- Main(and very important) difference between main stream AI (Good old fashion AI, knowledge-based systems) and the vast majority of machine learning (e.g.,Neural networks is:
  - Main stream AI is knowledge-driven while majority of ML is data-driven
  - ML uses "labelled" data points (many of them) as examples and the system generates a model of this data, and then use this model to tell something about a new data (representing for example a person applying for a job)
  - This type of ML systems are called "black box"

ML-pipeline

The "reasoning" processes underlying machine learning systems.



Why/when AI-fairness is important?

Many things become automated by machine learning:

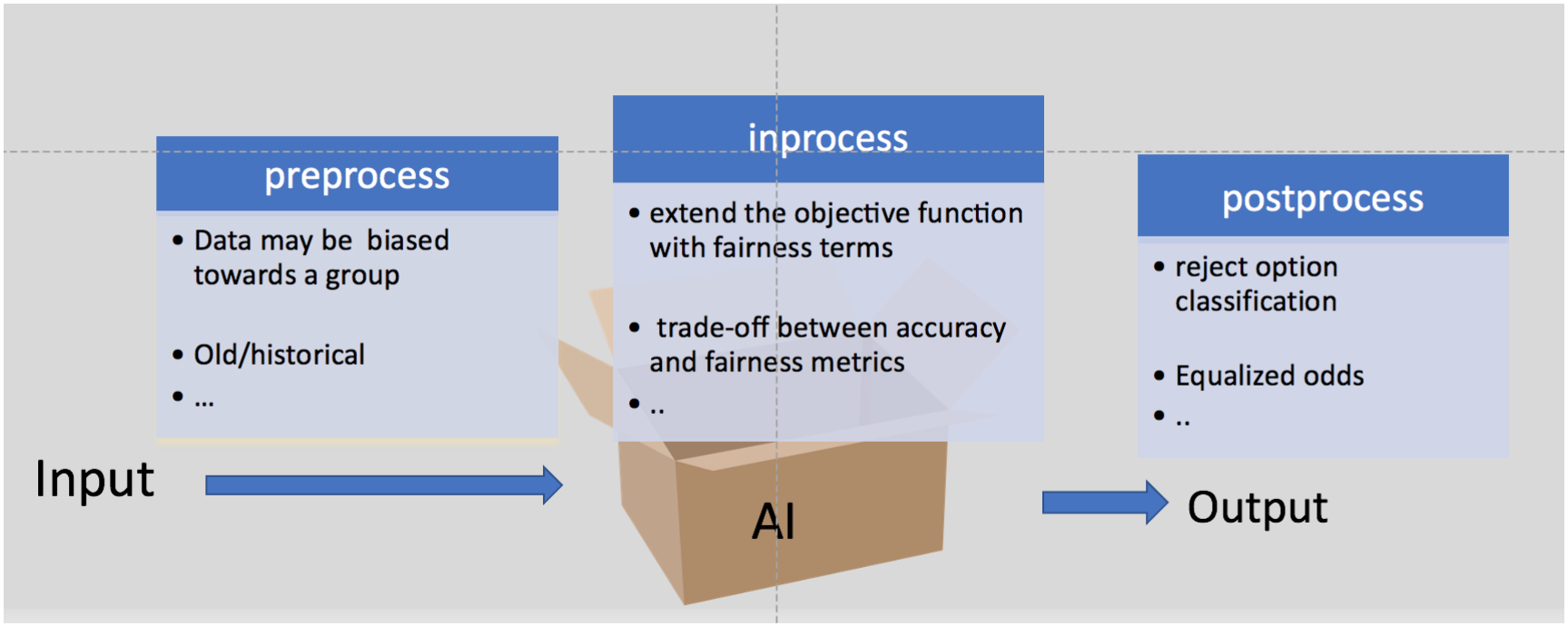
- employers select candidates by using by ML systems,
- Linked-In and XING use ML systems to rank candidates,
- courts in the US use ML systems to predict recidivism,
- banks use credit rating systems, which use ML,
- Amazon and Netflix use recommender systems
- If these system act unfair, groups and individuals may suffer.

Why is AI-fairness is important?

- Discrimination in a social sense of the word is prejudiced **treatment** of people based on perceived membership in certain classes, groups or categories, often called *protected classes*. The attribute that defines a protected class is called a *sensitive attribute.*, e.g., gender, race, religion, disability, or age.
- Unfairness is to limit people's life chances based not on merits but based on sensitive attributes like gender or race.
- In some cases it is unintended discrimination, where different groups receive different outcomes or treatment even though their protected class membership was not explicitly considered in the decision process. This is called **disparate impact**.

Unfairness in ML

Where is the problem?
Where to look at?
How to measure it?
How to solve it?



Sources of Bias in Data

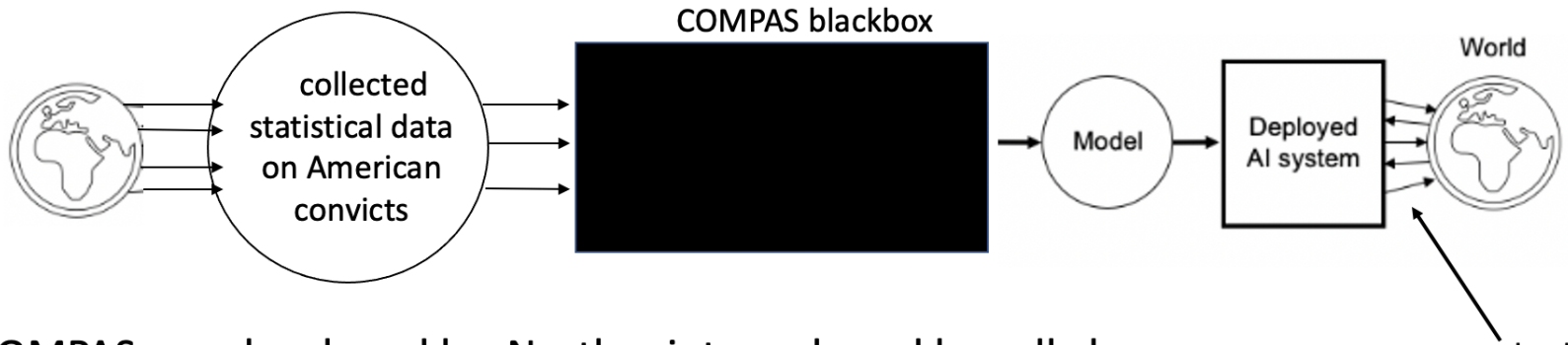
Bias type	Example	Reason
Under-coverage bias	Darker-skinned females being misclassified by facial analysis algorithms with error rates up to 34.7%, while the maximum error rate for lighter-skinned males is 0.8% [85].	Under-representation of dark-skinned females in datasets.
Measurement bias	Continuing on the example above, about darker-skinned individuals being misclassified by facial regcognition software.	Color balance settings and the dynamic range of cameras make it harder to capture high quality photos of people with dark skin than of people with light skin [86][61].
Label bias	The COMPAS crime recidivism tool has been shown to be biased against blacks in certain respects [3].	Crime recidivism prediction tools use future *arrests* as a proxy for future *crime*.
Historical bias / stale data	A Google Images search for "C.E.O." produced 11 percent women at a time when 27 percent of United States chief executives were women [5].	Historically, a lot less than 27% of CEOs were women, so if the search algorithm operated on historical data, it did not reflect current reality.
Aggregation bias	The COMPAS recidivism prediction tool has been shown to misclassify women as higher risk for violent crimes than they really are [63].	There are true differences between the genders in the sense that men are more likely to commit future violent crimes than women with the same criminal history [63].

Table 3.1: Some simple bias examples

[1] Gunnar Strand, MSc Thesis, IDI, NTNU, 2020

COMPAS example to Fairness

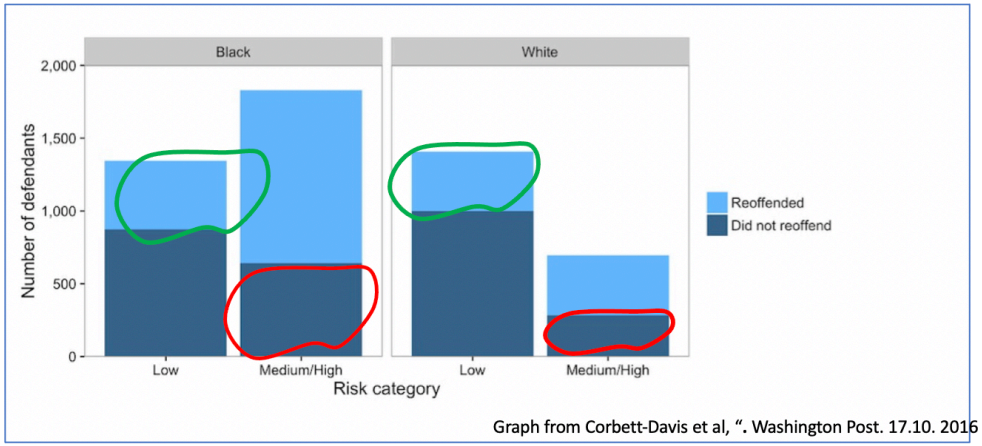
- This example is a very famous one creating hot debates in Fairness research and practice.
- It is about a system called COMPAS designed for predicting which criminals will reoffend.
- It is used by judges for risk estimation, across USA.



- COMPAS was developed by Northpointe and used broadly by e.g., judges in USA
- ”Recidivism Risk score" indicates how likely the convict would reoffend if released
- In 2016: Propublica organisation claimed that it was discriminatory (unfair) against black people

COMPAS example to Fairness - 2

ProPublica argues: COMPAS is not fair



Graph from Corbett-Davis et al, ". Washington Post. 17.10. 2016

	predicted high risk	predicted low risk
reoffended	True. Positive.	False Negative
not reoffended	False. Positive.	True Negative

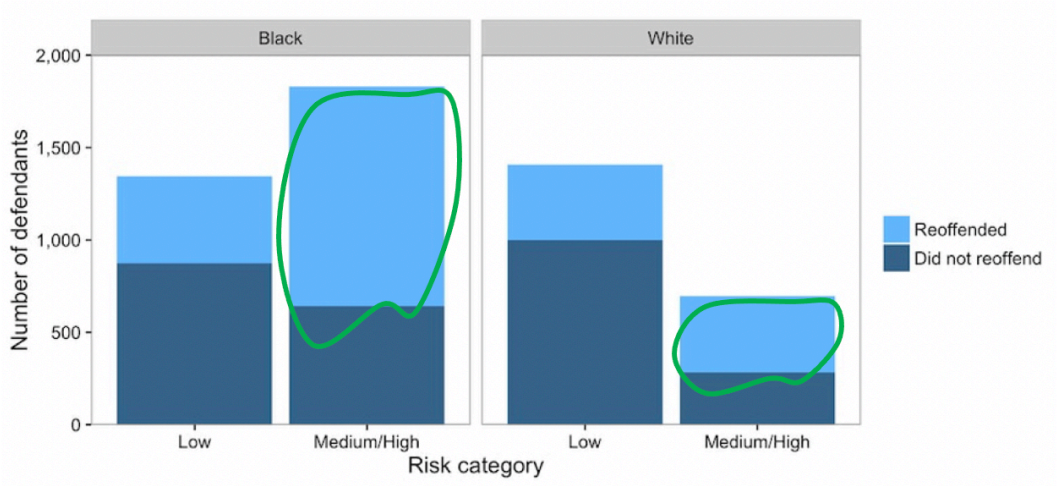
- COMPAS is unfair because likelihood of a non-reoffending black defendant to be predicted as high risk is twice of white defendant.
- Focus is on False positive and False negative rates

COMPAS example to Fairness - 3

Northpointe argues: COMPAS is fair

- COMPAS is fair because: among defendants that COMPAS predicted as “highly probable to reoffend”:
  - 60 % of the white defendants actually did reoffended
  - 61 % of the black defendents reoffended

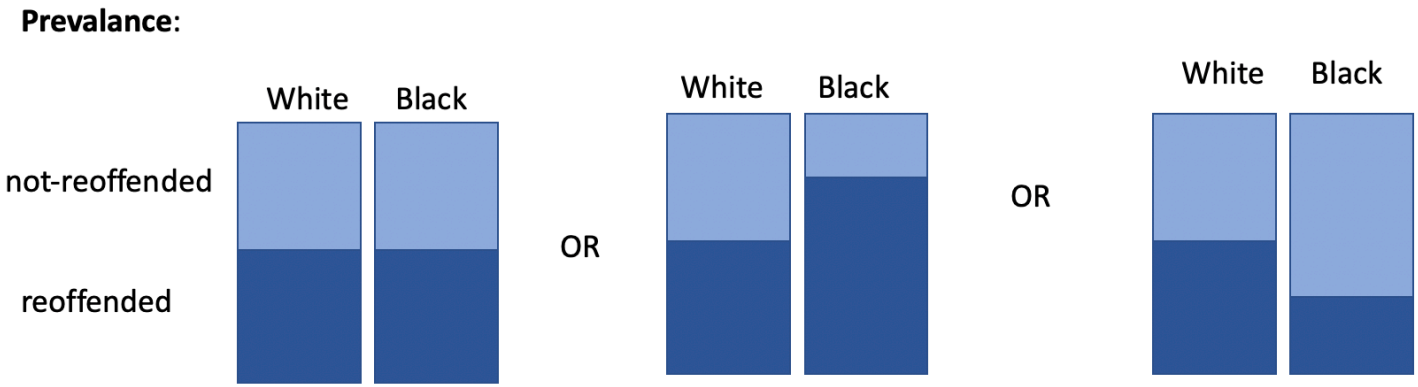
- Focus is more on the True positive (“predictive parity”)



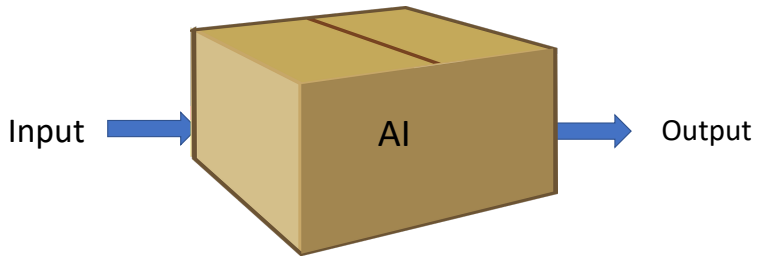
COMPAS example to Fairness - 4

Are both fairness goals possible simultaneously?

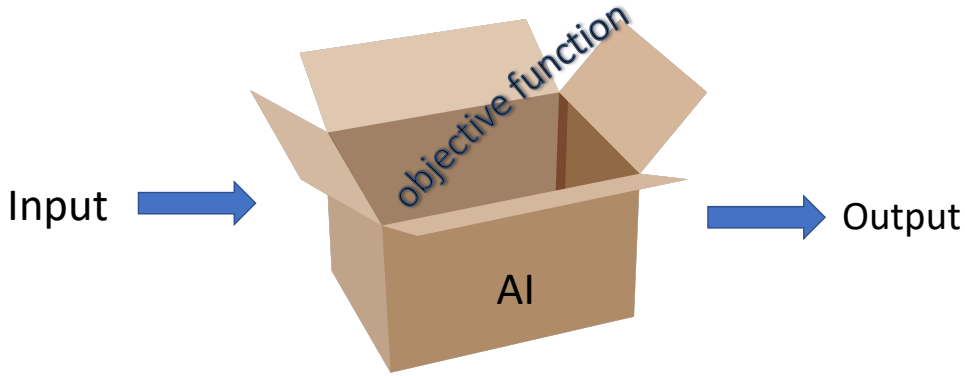
- Not always. It is not possible when the recidivism *prevalance* differs across groups.
- In COMPAS data, prevalance was different, around 52% vs 39% (according to Corbett-Davis et al analysis)



Performance of an AI system



- In this course we talked about PEAS (i.e. performance, environment, actuators, sensors), and defined rational behaviour as
  - *an agent's taking actions that maximize the expected value of the* **performance measure** *given the percept sequence so far.*
- Performance of an agent is measured through its "objective function"



Objective Function

- can be defined either as cost/loss function (to be minimized) or utility/profit/fitness function (to be maximized)
- and can be evaluated to a value, e.g., 90% of the dirt is sucked, fruit classes are predicted with 70% of accuracy.

Traditionally, objective function does not include ethical considerations such as fairness, privacy etc.

- . ...And adding "ethical objective aspects" to the objective function may interfere with the measures of traditional objective function
- Often there is a tradeoff between for example accuracy and fairness measures.

Perspectives on Ethics In Philosophy

- Different perspectives in Ethics in Philosophy:
- Deontological
- Consequentialist, Utilitarian
- Virtue

Deontological vs Consequentialism

From "Stanford Encyclopedia of Philosophy":

- In contemporary moral philosophy, deontology is one of those kinds of normative theories regarding which choices are morally required, forbidden, or permitted. In other words,
- deontology falls within the domain of moral theories that guide and assess *our choices* of what we ought to do (deontic theories), in contrast to those that guide and assess *what kind of person we are* and should be ("virtue" theories).

Deontological approach

- Deontologist's claim: some actions have inherent moral value – as required, forbidden, etc.
- Whether an act is morally right or wrong depends on whether it is in conformity or conflict with moral duties and rights.
- Moral principles and rules.

Consequentialist approach (from Stanford Encyclopedia)

Consequentialists hold that choices—acts and/or intentions—are to be morally assessed solely by the states of affairs they bring about. Consequentialists thus must specify initially the states of affairs that are intrinsically valuable—often called, collectively, "the Good." They then are in a position to assert that whatever choices increase the Good, that is, bring about more of it, are the choices that it is morally right to make and to execute. (The Good in that sense is said to be prior to "the Right.")

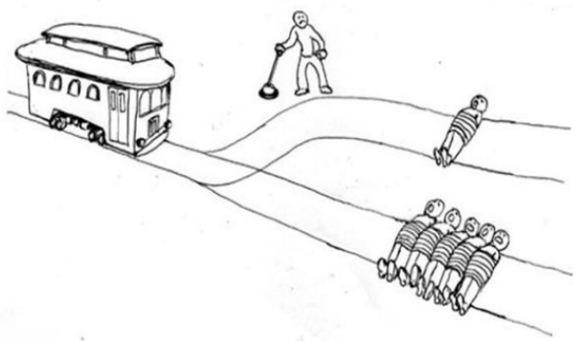
Consequentialists can and do differ widely in terms of specifying the Good. Some consequentialists are monists about the Good. Utilitarians, for example, identify the Good with pleasure, happiness, desire satisfaction, or "welfare" in some other sense. Other consequentialists are pluralists regarding the Good. Some of such pluralists believe that how the Good is distributed among persons (or all sentient beings) is itself partly constitutive of the Good, whereas conventional utilitarians merely add or average each person's share of the Good to achieve the Good's maximization.

Moral Machines?

- Philosophers usually consider machines as not capable of making moral decisions.
- However, one can try to find properties such that machines could act. morally.
- Machines need to have at least
  - beliefs about the world
  - intentions
  - moral knowledge
  - the possibility to compute what consequences ones own action can have,in which case they can be considered as moral agents.

Famous "Trolley Problem" and Autonomous driving

From Wiki: The trolley problem is a series of thought experiments in ethics and psychology, involving stylized ethical dilemmas of whether to sacrifice one person to save a larger number.



- The story goes like this: There is a runaway trolley barrelling 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 on the side track. You have two options:
  - Do nothing and allow the trolley to kill the five people on the main track.
  - Pull the lever, diverting the trolley onto the side track where it will kill one person.

  Which is the more ethical option? Or, more simply: What is the right thing to do?

Famous "Trolley Problem" and Autonomous driving-cont.

Trolley problems highlight the difference between deontological and consequentialist ethical systems The central question that these dilemmas bring to light is on whether or not it is right to actively inhibit the utility of an individual if doing so produces a greater utility for other individuals.

Click on the link below for a video illustrating Trolley-like dilemma in autonomous driving:
Ethical issues regarding Autonomous Driving

Summary

- If a new AI winter arises, it will be due to ethical problems
- What are the main ethical concerns
- Several AI-ethics Regulations
- Philosophical approaches to ethics and morality
- AI is not equal to Machine Learning
- Why and how unfairness (and discrimination) occur in ML?
- Fairness definitions and metrics - some
- About mitigation of fairness
- Trolley problem and the link to Consequentialism (in Philosphy of ethics)

Acknowledgement

This slide set includes some slides or parts of slides from

- Joschka Boedecker et al.