Ethical Implications of AI
- series of lectures-

coordinator
Prof. Roberto V. Zicari
Frankfurt Big Data Lab
www.bigdata.uni-frankfurt.de

April 22 till June 18, 2020

Spring

Artificial Intelligence (AI) seems the defining technology of our time.

John McCarthy defines AI, back in 1956 like this:

"*AI involves machines that can perform tasks that are characteristic of human intelligence*".

John McCarthy **--1956**

Artificial Intelligence (AI)

What are the main differences between:

- Artificial Intelligence,
- Machine Learning and
- Deep Learning

?

Machine Learning

To put it simply, Machine Learning is a way of achieving AI.

- Arthur Samuel's definition of Machine Learning (ML) is from **1959**:

"Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed".

Typical problems solved by Machine Learning

Typical problems solved by Machine Learning are:

- Regression.
- Classification.
- Segmentation.
- Network analysis.

Big Data

- What has changed dramatically since those pioneering days is the rise of **Big Data** and of **computing power**, making it possible to analyze massive amounts of data at scale!

Deep Learning

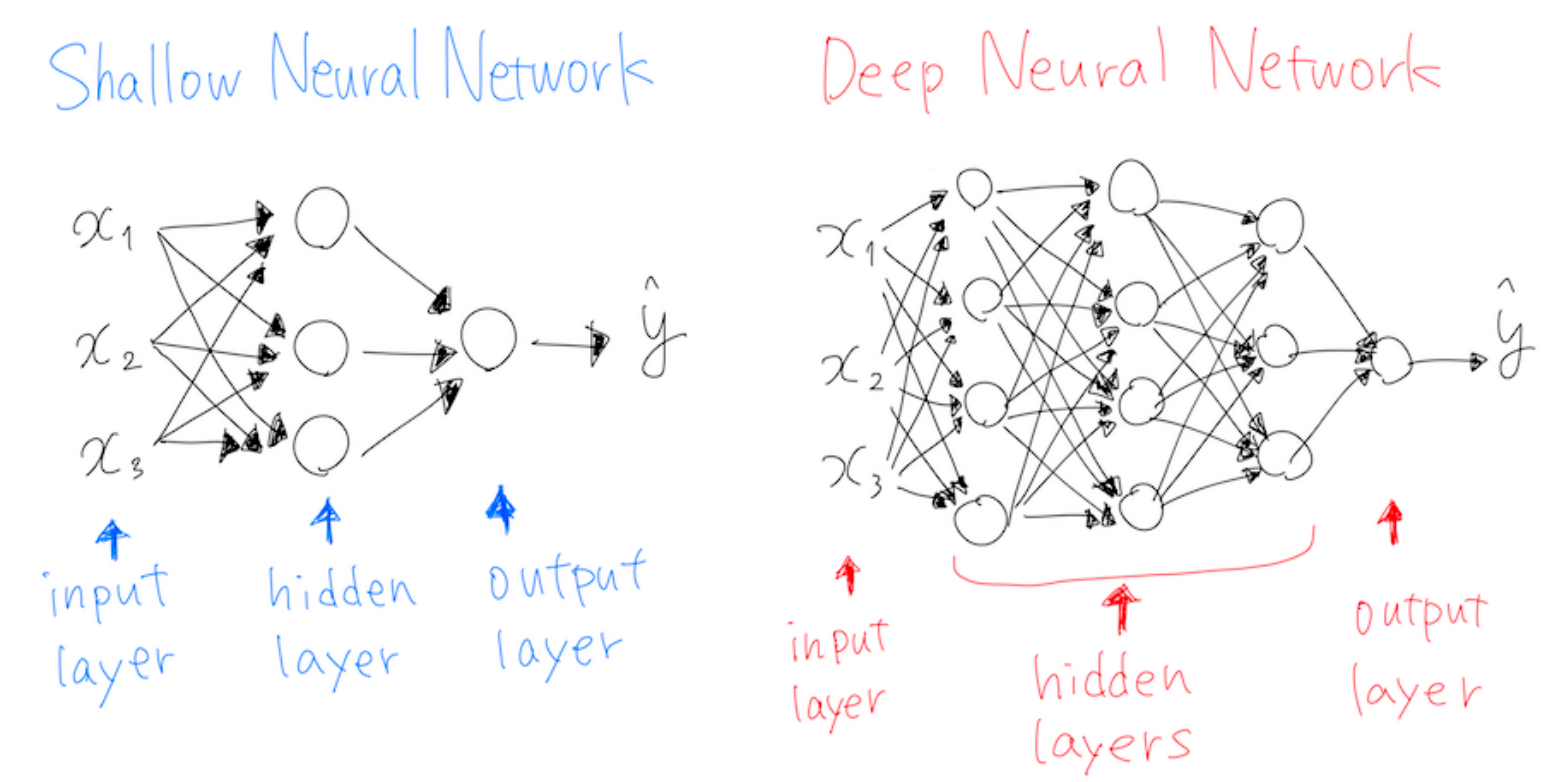
- AI needs Big Data and Machine Learning to *scale*. Machine learning is a way of *"training"* an algorithm so that it can learn.
- Huge amounts of data are used to train algorithms and allowing algorithms to "learn" and improve.
- Deep Learning is a subset of Machine Learning and was inspired by the structure and function of the brain.

Neural Networks

- Example:

Artificial Neural Networks(ANNs), are algorithms that resemble the biological structure of the brain, namely the interconnecting of many neurons

Shallow vs. Deep Neural Networks



- A Comparison of Shallow vs. Deep Neural Networks (source: BecomingHuman.ai)

Use of Deep Learning

Deep learning architectures such as deep neural networks, deep belief networks, recurrent neural networks and convolutional neural networks
have been applied to fields including:

computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game programs,

where they have produced results comparable to and in some cases superior to human experts.

Source: Wikipedia

Convolutional Neural Networks

- (**ConvNets** or **CNNs**) are a category of Neural Networks that have proven very effective in areas such as image recognition and classification. ConvNets have been successful in identifying faces, objects and traffic signs apart from powering vision in robots and self driving cars.

Image Recognition Classifier

Examples of Tools and Technologies

Anaconda —free and open source distribution of the Python and R programming languages for data science and machine learning related applications, that aims to simplify package management and deployment. *https://www.anaconda.com/download/*

Spyder open source cross-platform IDE for scientific programming in the Python language. It comes installed with anaconda

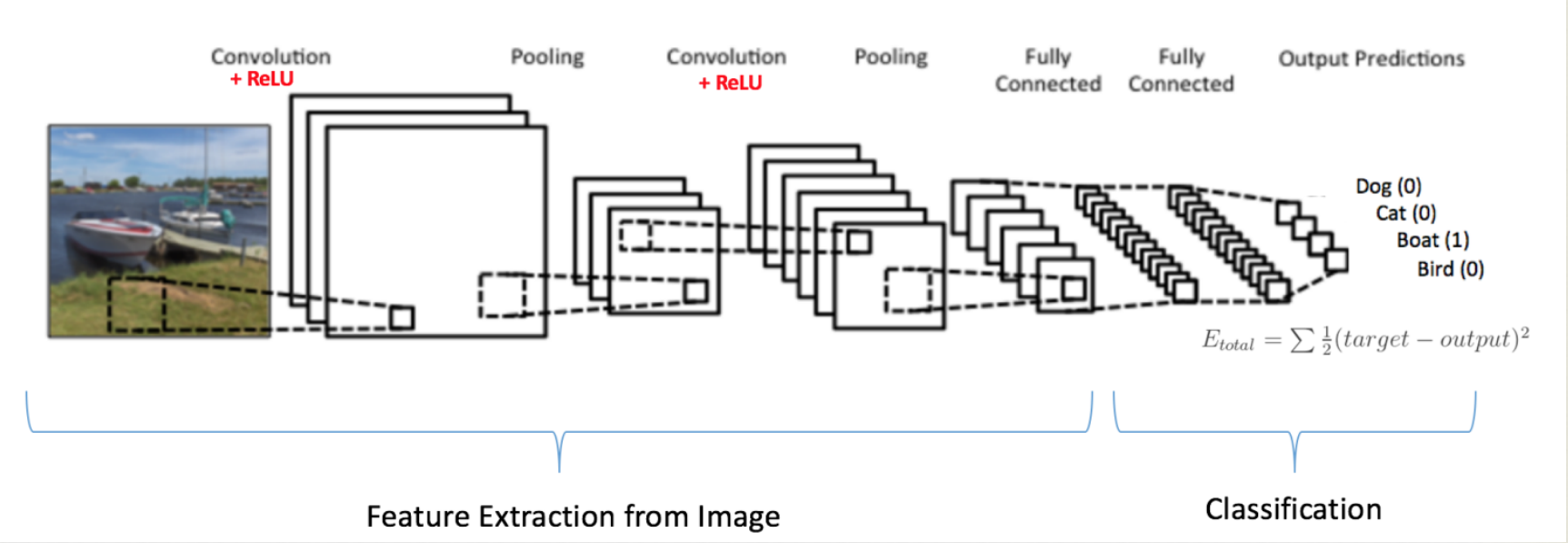
Tensorflow —open-source software library for dataflow programming across a range of tasks. *https://www.tensorflow.org/install/install_windows*

Keras —open source neural network library written in Python.

CNN — Convolution Neural network , a class of deep, feed-forward artificial **neural networks**, most commonly applied to analyzing visual imagery.

Source: https://medium.com/nybles/create-your-first-image-recognition-classifier-using-cnn-keras-and-tensorflow-backend-6eaab98d14dd

Training a ConvNet: Output predictions: Probability.



Training ConvNet
source:
https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

AI and Society

- With AI the important question is how to avoid that it goes out of control, and how to understand how decisions are made and what are the consequences for society at large.

AI, Ethics, Law

- AI is becoming a sophisticated tool in the hands of a variety of stakeholders, including political leaders. Some AI applications may raise new ethical and legal questions, and in general have a significant impact on society (for the good or for the bad or for both).

Ethical implications of AI: Key Questions:

- What are the consequences for society?
- For human beings / individuals?
- Does AI serve human kind?

Ethical reflection

- Discussion and debate of ethical issues is an essential part of professional development—both within and between disciplines—as it can establish a mature community of responsible practitioners.

- Through ethical reflection one can gain orientation / competencies that will help in the ethical decision making.

Our view of the world

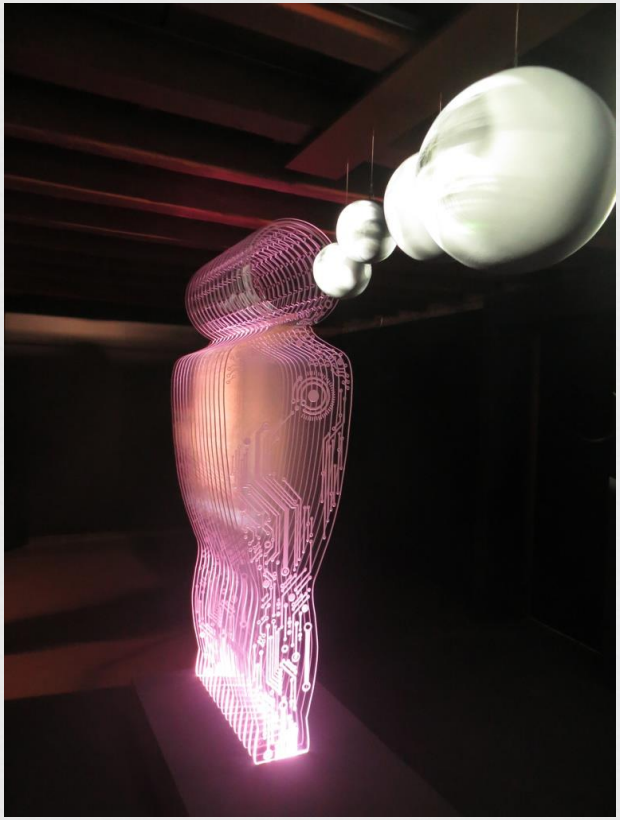
Contemporary Western European democracy.

Fundamental values

The essence of a modern democracy is based on respect for others, expressed through support for fundamental human rights.

- Source:
Ethical Business Regulation:Understanding the Evidence , Christopher Hodges
Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford February 2016

Findings of behavioural psychology



How people make decisions: fairness, proportionality, morality

"**Decisions are made by people** rather than by organisations, although the structures, systems, objectives, culture and incentives that operate within organisations can affect the decisions made by the people who work in them.

Empirical research has found that people obey rules where:

- the rule corresponds to their internal moral value system;
- the rule has been made fairly; and
- the rule is applied fairly. "

Source:
Ethical Business Regulation:Understanding the Evidence , Christopher Hodges
Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford February 2016

Broad sense of justice

- „Individuals do not need to be familiar with every legal rule, but generally operate on a *broad sense of justice*. When faced with a particular situation, the individual may draw on examples of a social norm or custom as a means of guiding social compliance and normative orientation.“

- Source:

Ethical Business Regulation:Understanding the Evidence , Christopher Hodges
Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford February 2016

Social influences

- "Human behaviour is strongly *responsive to social influences*: people want to conform to the perceived behaviour of other people, and that influence can overcome known facts or one's own ideological worldview.

- Source:

Ethical Business Regulation:Understanding the Evidence , Christopher Hodges
Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford February 2016

Information, loss aversion

- "People can be influenced by *how information is presented or 'framed'*, so that information that is vivid and salient can have a larger impact on behaviour than information that is statistical and abstract.

- People often display *loss aversion*: they may well dislike losses more than they like corresponding gains. "

- Source:

Ethical Business Regulation:Understanding the Evidence , Christopher Hodges
Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford February 2016

Assessing probability

- People have *difficulties in assessing probability*:

people often show unrealistic optimism, may neglect or disregard the issue of probability, especially when strong emotions are triggered, and when emotions are strongly felt, may focus on the outcome and not on the probability that it will occur.

Judgments about probability are often affected by whether a recent event comes readily to mind.

- Source:

Ethical Business Regulation:Understanding the Evidence , Christopher Hodges
Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford February 2016

Respect for rules

- People's respect for all rules and for the system generally will be undermined where they see that rules are not being enforced evenly and fairly.

- Source:

Ethical Business Regulation:Understanding the Evidence , Christopher Hodges
Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford February 2016

Ethical business

"How can businesses behave ethically?

- The requirement is for a business to adopt ethical business practices in everything that is done throughout the organisation. Codes on individual aspects, such as production, waste, marketing or social responsibility, are not enough: the approach has to be **holistic**. It has to be led from the top, but to exist at every level of the social groups within an organisation. "

- Source:

Ethical Business Regulation:Understanding the Evidence , Christopher Hodges
Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford February 2016

Core values

- "Studies on the causes of sustained long term business success have concluded that it is critical to establish clear ***core values***, which are shared by ***all*** members of the workforce, form an ideology that is enduring and able to be applied consistently in different trading and geographical circumstances, whilst operational goals are constantly examined and developed. "

- Source:

Ethical Business Regulation:Understanding the Evidence , Christopher Hodges

Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford February 2016

Evidence of trust

- "It is essential to provide *evidence of trust* that an organisation operates with ethical values, to support independent judgment on whether an expectation of ethical behaviour is warranted.

- Mere claims by a company that it can be trusted will clearly not suffice. *Mechanisms should be designed to produce reliable evidence of trust*. "

- Source:

Ethical Business Regulation:Understanding the Evidence , Christopher Hodges
Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford February 2016

Response to adverse events

- "In most cases, where people are acting in good faith, the appropriate response to adverse events is to support them to analyse and learn rather than to blame. Failures should be noted and acknowledged, rather than ignored, and an appropriate response made. "

- Source:

Ethical Business Regulation:Understanding the Evidence , Christopher Hodges
Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford February 2016

Response to adverse events

- "Where actions are immoral, or accountability as described above has not been observed, a proportionate response should be made. Enforcement policies should generally avoid the concept of deterrence, since it has limited effect on behaviour, conflicts with a learning-based performance culture, and is undemocratic. "

- Source:

Ethical Business Regulation:Understanding the Evidence , Christopher Hodges
Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford February 2016

Response to adverse events

- "Where sanctions are imposed, the totality of the sanctions should be proportionate to the degree of moral culpability involved. That requires an equalisation as between all of the various factors: social response, reputational and public response, employment discipline response, civil redress response, and regulatory or criminal response. "

- Source:

Ethical Business Regulation:Understanding the Evidence , Christopher Hodges
Professor of Justice Systems, and Fellow of Wolfson College, University of Oxford February 2016

Benefits & Risks of Artificial Intelligence

AI and Big Nudging

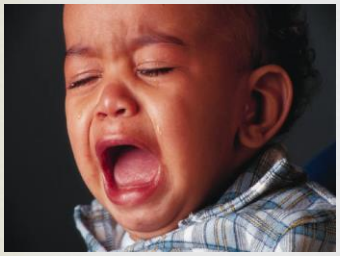
"Big Nudging"

He who has large amounts of data can manipulate people in subtle ways. But even benevolent decision-makers may do more wrong than right.

- **Spotlight on China: Is this what the Future of Society looks like?**

How would *behavioural* and *social control* impact our lives? The concept of a Citizen Score, which is now being implemented in China, gives an idea.

Source: Will Democracy Survive Big Data and Artificial Intelligence?. Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoven, J., Zicari, R. V., & Zwitter, A.. (2017). Scientific American (February 25, 2017).



Do no harm
Can we explain decisions?

What if the decision made using AI-driven algorithm harmed somebody, and you cannot explain how the decision was made?

- At present we do not really understand how Advanced AI-techniques such as used in **Deep learning** (e.g. **neural networks**) really works. It can be extremely difficult to understand which features (millions of features) of the data the machine used, and how they were weighted, to contribute to the outcome.
- This is due to the technical complexity of such advanced neural networks, which need huge amount of data to learn properly. It is a try and error.
- This poses an ethical and societal problem.

Bias in machine learning

Technically *(*)* Bias in machine learning= errors in estimation or over/under representing populations when sampling.

Selection, sampling, reporting bias

Bias of an estimator

Inductive bias

(*) Source: CS 294: Fairness in Machine Learning UC Berkeley, Fall 2017 https://mrtz.org/nips17/#/6

Understanding Bias in Algorithmic Design

" Suppose two people are tasked with developing a system to sort a basket of fruit.

They have to determine which pieces are "*high quality*" and will be sold at the market, and which will instead be used for making jam. "

Source: Understanding Bias in Algorithmic Design
https://medium.com/impact-engineered/understanding-bias-in-algorithmic-design-db9847103b6e

Understanding Bias in Algorithmic Design

Both people are given the exact same data — the fruit — and the same task... Given the same task and data, the two people are likely to have different results.

Source: Understanding Bias in Algorithmic Design
https://medium.com/impact-engineered/understanding-bias-in-algorithmic-design-db9847103b6e

Understanding Bias in Algorithmic Design

Perhaps one person believes the primary indicator of a fruit's quality is *brightness of color*. That person may sort the fruit based on how vibrant it is, even though not all fruits are brightly colored; **that person would send strawberries to the market and melons to the jam factory.**

Meanwhile, the other person might believe that *unblemished* fruit is the best quality, even though fruits with protective rinds might look scruffy on the outside, but are perfectly fine on the inside; **that person could send unripe strawberries to the market and ripe melons or bananas to the jam factory.**

Source: Understanding Bias in Algorithmic Design
https://medium.com/impact-engineered/understanding-bias-in-algorithmic-design-db9847103b6e

ALGORITHMIC CONSEQUENCES

Similarly logical and evenly applied criteria, will result in two different outcomes for the same basket of fruit.

It's one thing to have an algorithm that marginalizes melons or unfairly sorts cucumbers, **but what happens when algorithms make important decisions about humans?**

Source: Understanding Bias in Algorithmic Design
https://medium.com/impact-engineered/understanding-bias-in-algorithmic-design-db9847103b6e

Bias and Discrimination

When algorithms are used for example, to review loan applications, recruit new employees or assess potential customers, **if the data are skewed** the decisions recommended by **such algorithms may be discriminatory against certain categories or groups.**

Other kinds of bias

Allocative harm= when a system allocates or withholds a certain opportunity or resource

Representation harm = when a system reinforces the subordination of some groups along the lines of identity

Source:
Kate Crawford, Keynote "The Trouble with Bias" Neural Information Processing System Conference

False Positive, False Negative

- A **false positive** is an error in data reporting in which a test result improperly indicates presence of a condition, such as a disease (the result is *positive*), when in reality it is not present,

- A **false negative** is an error in which a test result improperly indicates no presence of a condition (the result is *negative*), when in reality it is present.

Source: Wikipedia

AI, Context and Levels of Harms

The overall potential damage that an AI systems may cause in its respective social process *depends on the context.*

Do you agree with this quote?

"What happens if my algorithm is wrong? Someone sees the wrong ad. What's the harm? It's not a ***false positive*** *for breast cancer." (*)*

-- ***Claudia Perlich,*** Data Scientist, 2016

(*) Source: **Big Data and The Great A.I. Awakening. Interview with Steve Lohr** ODBMS Industry Watch, December 19, 2016

Concept Building

An important obstacle to progress on the ethical and societal issues raised by AI-based systems is the *ambiguity* of many central *concepts* currently used to identify salient issues:

- **Terminological overlaps**
- **Differences between disciplines**
- **Differences across cultures and publics**

Source:[1] *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Terminological overlaps: Bias, Fairness , and Discrimination

The terms '**bias**', '**fairness**', and '**discrimination**' are often used to refer to problems involving datasets or algorithms which (in some sense) disadvantage certain individuals or groups.

It is not clear that all cases referred to by these terms involve the same type of problem.

Source:[1] *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research*. Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Terminological overlaps : **Discrimination**

Barocas (2014) distinguishes three kinds of *concerns* for algorithms based on data-mining, which have been raised under the heading '**discrimination**'.

SourceBarocas, S. (2014). Data mining and the discourse on discrimination. Proceedings of the Data Ethics Workshop, Conference on Knowledge Discovery and Data Mining (KDD).

Terminological overlaps : **Discrimination**

1. Cases where deployers of an algorithm *deliberately* attempt to disadvantage certain users and make this difficult to detect (e.g. by hiding the critical bit of code within a complicated algorithm).

Source Barocas, S. (2014). Data mining and the discourse on discrimination. Proceedings of the Data Ethics Workshop, Conference on Knowledge Discovery and Data Mining (KDD).

Terminological overlaps : **Discrimination**

2. Cases where data-mining techniques *produce errors* which disadvantage certain users (e.g. due to unreliable input data or users drawing faulty inferences from the algorithms' output).

Terminological overlaps : **Discrimination**

3. Cases where an algorithm *enhances decision-makers' ability to distinguish and make differential decisions between people* (e.g. allowing them to more accurately identify and target financially vulnerable individuals for further exploitation).

Differences between disciplines
Bias

- For example, in *statistics* a '**biased sample**' means a sample that does not adequately represent the distribution of features in the reference population (e.g. it contains a higher proportion of young men than in the overall population).

- Source:[1] *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Differences between disciplines
Bias

- In *law* and *social psychology*, by contrast, the term '**bias**' often carries the connotation of negative attitudes or prejudices towards a particular group.

- In this sense, a dataset which is 'unbiased' (in the statistical sense) may nonetheless encode common biases (in the social sense) towards certain individuals or social groups.

Source:[1] *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Differences between disciplines

Avoiding cross-talk

- Distinguishing these different uses of the same term is important to avoid cross-talk.

Differences across cultures and publics

- In addition to disciplinary differences, key concepts may be understood differently or carry different connotations across different *cultures* and *publics* (*).

(*) use the term '*publics*' in plural to emphasise that different interest groups (scientists, mediators, decision-makers, activists, etc.) bring their own distinct perspectives.

Source:[1] *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Differences across cultures and publics
Privacy

Example: Concept of **privacy**.

- Western ethical traditions (e.g. Kantianism) tend to conceive of individual privacy as an intrinsic good

Source:[1] *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Differences across cultures and publics Privacy

Example: Concept of **privacy**.

This is often not the case in Eastern traditions:

- "In Confucianism, which tends to emphasise the collective good over the individual, the notion of individual privacy (as opposed to the collective privacy e.g. of a family) has traditionally been given less attention (and may even carry negative connotations e.g. of shameful secrets). "

Source:[1] *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Differences across cultures and publics Privacy

Example: Concept of **privacy**.

- "Traditional Buddhism regards the belief in an autonomous self as a pernicious illusion, some Buddhist traditions have argued one should actively share one's secrets as a means to achieving a lack of self."

Source:[1] *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research*. Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Too homogeneous? Diversity

"If AI/ML teams are too *homogeneous*, the likelihood of group-think and one-dimensional perspectives rises – thereby increasing the risk of leaving the whole AI/ML project vulnerable to *inherent biases and unwanted discrimination*."

-- Nicolai Pogadl (*)

(*) Source: personal communication.

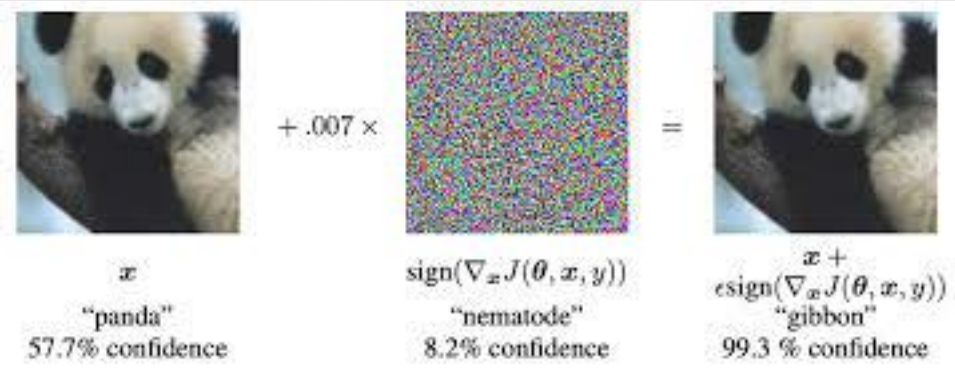
AI Safety

- "Deep neural networks can fail to generalize to **out-of-distribution inputs**, including *natural, non-adversarial ones*, which are common in real-time settings". (*)

(*) Source : **Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects** Michael A. Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, Anh Nguyen *(Submitted on 28 Nov 2018 (v1), last revised 13 Jan 2019 version, v2*

AI-attacks

- "Several machine learning models, including neural networks, consistently misclassify **adversarial examples**---inputs formed by applying small but intentionally worst-case perturbations to examples from the dataset, such that the perturbed input results in the model outputting an incorrect answer with high confidence."



(**) **Source: Explaining and Harnessing Adversarial Examples** Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy *(Submitted on 20 Dec 2014 (v1), last revised 20 Mar 2015 version, v3)*

Example : Autonomous Cars

Let`s consider an autonomous car that relies entirely on an algorithm that had taught itself to drive by watching a human do it.

What if one day the car crashed into a tree, or even worse killed a pedestrian?

The Uber Case for ***False positive*** *for plastic bags…*

„The newsletter "The Information" has reported a leak from Uber about their fatal accident.

The relevant quote:

(*) How reliable is this Source? : https://ideas.4brad.com/uber-reported-have-made-error-tuning-perception-system

Story also in Der Spiegel Nr. 50/8.12.2018 *Tod durch Algorithms* (Philipp Oehmke)

The Uber Case

"The car's sensors detected the pedestrian, who was crossing the street with a bicycle, but Uber's software decided it didn't need to react right away. **That's a result of how the software was tuned.**

Like other autonomous vehicle systems, Uber's software has the **ability to ignore "false positives**," or objects in its path that wouldn't actually be a problem for the vehicle, such as a plastic bag floating over a road.

In this case, Uber executives believe the company's **system was tuned so that it reacted less to such objects.** But the tuning went too far, and the car didn't react fast enough, one of these people said." (*)

Algorithms learn from data

"Since the algorithms learn from data, it's not as easy to understand what they do as it would be if they were programmed by us, like traditional algorithms.

But that's the essence of machine learning: that it can go beyond our knowledge to discover new things.

A phenomenon may be more complex than a human can understand, but not more complex than a computer can understand. (*)

--- Pedro Domingos

*) Source: **On Artificial Intelligence, Machine Learning, and Deep Learning. Interview with Pedro Domingos**, ODBMS Industry Watch, June 18, 2018

Algorithms learn from data

"And in many cases we also don't know what humans do: for example, we know how to drive a car, but we don't know how to program a car to drive itself.

But with machine learning the car can learn to drive by watching video of humans drive." (*)

--- Pedro Domingos

(*) Source: **On Artificial Intelligence, Machine Learning, and Deep Learning. Interview with Pedro Domingos**, ODBMS Industry Watch, June 18, 2018

Machine Learning Limits: Causality

- "**Causality**—in other words, grasping not just patterns in data but ***why*** something happens. Why is that important, and why is it so hard?

- " If you have a good causal model of the world you are dealing with, you can generalize even in ***unfamiliar*** situations. That's crucial. We humans are able to project ourselves into situations that are very different from our day-to-day experience.
- **Machines are not, because they don't have these causal models**."

--Yoshua Bengio

(*) Source MIT Technology Review
https://www.technologyreview.com/s/612434/one-of-the-fathers-of-ai-is-worried-about-its-future/

Machine Learning Limits: Causality

- "We can hand-craft them, but that's not enough. We need machines that can discover causal models. To some extent it's never going to be perfect. We don't have a perfect causal model of the reality; that's why we make a lot of mistakes. But we are much better off at doing this than other animals.

- **Right now, we don't really have good algorithms for this**, but I think if enough people work at it and consider it important, we will make advances."

  --Yoshua Bengio (*) Source MIT Technology Review

Use Case Waymo Self Driving Cars

Waymo (a subsidiary of Alphabet Inc) created a Recurrent Neural Network (RNN) for Driving.

They trained the neural network Imitating the "**Good**" and synthesizing the "**Bad**".

„They trained the model with examples from the equivalent of about 60 days of **expert driving** data, while including training techniques such as past motion dropout to ensure that the network doesn't simply continue to extrapolate from its past motion and actually responds correctly to the environment."

(*) Source : **Learning to Drive: Beyond Pure Imitation** Dec 10, 2018, https://medium.com/waymo/learning-to-drive-beyond-pure-imitation-465499f8bcb2

ChauffeurNet: Learning to Drive by Imitating the Best and Synthesizing the Worst. Mayank Bansal, Alex Krizhevsky, Abhijit Ogale, Dec 7, 2018 https://arxiv.org/pdf/1812.03079.pdf

Use Case Waymo
Learning from "Bad Examples"

"It's not difficult to feed the bad examples. That's what we do in our training, we feed it synthesized bad examples and add a training loss that tells the network not to emulate the bad behavior.

Real examples of bad behavior are difficult to intentionally obtain, and it is **simpler** and **safer** to **synthetically** create bad examples in **simulation**."
--Abhijit Ogale

- *Source: Personal communication*

Use Case Waymo
The WHY question

”Knowing **why** an expert driver behaved the way they did and what they were reacting to is critical to building a causal model of driving. For this reason, simply having a large number of expert demonstrations to imitate is not enough.

Understanding the why makes it easier to know how to improve such a system, which is particularly important for safety-critical applications.” (*)

However, I do not believe that we know WHY and HOW we drive though…

Try for yourselves: Explain to another person how do you drive and why you react in certain situations they way you do….. And please let me know the result.

(*) Source : **Learning to Drive: Beyond Pure Imitation**
https://medium.com/waymo/learning-to-drive-beyond-pure-imitation-465499f8bcb2

Use Case Waymo
Intelligence and Ethical Decisions

- "As a layperson looking at this particular field of ethical systems, I see some parallels between determining whether a system has intelligence and whether a system is making ethical decisions or not.
- In both cases, we are faced with a **kind of Turing test scenario** where we find it difficult to articulate what we mean by intelligence or ethics, and can only probe a system in a Turing test manner to determine that it is indistinguishable from a model human being."

--- **Abhijit Ogale**

Disclaimer: personal viewpoint as a ML researcher, not in his role at Waymo.

- *Source: Personal communication*

Use Case Waymo
How ML systems learns

- "The trouble with this approach though is that we are assuming that if the system passes the test, it shares the same or similar internal representations as the human tester, and it is likely that its intelligence or ethical behavior generalizes well to new situations. We do the same to assess whether another human is ethical or not.

- **This is a great difficulty, because we currently know that our artificial ML systems learn and generalize differently than humans do, so this kind of approach is unlikely to guarantee generally intelligent or ethical behavior. "**

- --- **Abhijit Ogale**
  Disclaimer: personal viewpoint as a ML researcher, not in his role at Waymo.

Use Case Waymo
Reduce likelihood of risks

- **" I think the best we can currently do is to explicitly engineer/bound and rigorously test the system against a battery of diverse scenarios to check its decisions and reduce the likelihood of undesirable behavior.**

- **The number of tests needs to be large and include long-tail scenarios because deep learning systems don't have as large a generalization horizon as human learning, as evidenced by their need of a mountain of training data. "**

--- **Abhijit Ogale**

Disclaimer: personal viewpoint as a ML researcher, not in his role at Waymo.

Learning: Who sets the examples?

"If the learning took place before the car was delivered to the customer, the car's manufacturer would be liable, just as with any other machinery.

The more interesting problem is if the car learned from its driver.

Did the driver set a bad example, or did the car not learn properly?"

--Pedro Domingos

(*) Source: **On Artificial Intelligence, Machine Learning, and Deep Learning. Interview with Pedro Domingos**, ODBMS Industry Watch, June 18, 2018

AI and Regulations

Policy Makers and AI

*"**Citizens** and **businesses** alike need to be able to **trust** the technology they interact with, and have effective safeguards protecting fundamental rights and freedoms.*

*In order to increase **transparency** and **minimise the risk of bias**, AI systems should be developed and deployed in a manner that allows humans to **understand** the basis of their actions.*

***Explainable AI** is an essential factor in the process of strengthening people's trust in such systems."* (*)

-- Roberto Viola

Director General of DG CONNECT (Directorate General of Communication Networks, Content and Technology) at the European Commission.

- *(*) Source* On the Future of AI in Europe. Interview with Roberto Viola, ODBMS Industry Watch2018-10-09

Trustworthy artificial intelligence.

EU High-Level Expert Group on AI presented their ethics guidelines for trustworthy artificial intelligence:

- (1) **lawful** - respecting all applicable laws and regulations
- (2) **ethical** - respecting ethical principles and values
- (3) **robust** - both from a technical perspective while taking into account its social environment

Ethical Principles in the Context of AI Systems

EU **four ethical principles**, rooted in fundamental rights

(i) Respect for human autonomy

(ii) Prevention of harm

(iii) Fairness

(iv) Explicability

- **Tensions between the principles**

e.g. Consider as an example the use of AI systems for 'predictive policing', which may help to reduce crime, but in ways that entail surveillance activities that impinge on individual liberty and privacy.

Ethical tensions

- "We use the umbrella term 'tension' to refer to different ways in which values can be in conflict, some more fundamentally than others.

- When we talk about tensions between values, we mean tensions between the pursuit of different values in technological applications rather than an abstract tension between the values themselves."

Source:[1] *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research*. Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

EXAMPLES OF TENSIONS BETWEEN VALUES

- **Quality of services *versus* privacy**: using personal data may improve public services by tailoring them based on personal characteristics or demographics, but compromise personal privacy because of high data demands.

- **Personalisation *versus* solidarity**: increasing personalisation of services and information may bring economic and individual benefits, but risks creating or furthering divisions and undermining community solidarity.

- Source:[1] *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

EXAMPLES OF TENSIONS BETWEEN VALUES

- **Convenience *versus* dignity**: increasing automation and quantification could make lives more convenient, but risks undermining those unquantifiable values and skills that constitute human dignity and individuality.

- **Privacy *versus* transparency**: the need to respect privacy or intellectual property may make it difficult to provide fully satisfying information about an algorithm or the data on which it was trained.

- Source:[1] *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research*. Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London*. Nuffield Foundation.

EXAMPLES OF TENSIONS BETWEEN VALUES

- **Accuracy *versus* explainability**: the most accurate algorithms may be based on complex methods (such as deep learning), the internal logic of which its developers or users do not fully understand.

- **Accuracy *versus* fairness**: an algorithm which is most accurate on average may systematically discriminate against a specific minority.

- Source:[1] *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London*. Nuffield Foundation.

EXAMPLES OF TENSIONS BETWEEN VALUES

- **Satisfaction of preferences *versus* equality**: automation and AI could invigorate industries and spearhead new technologies, but also exacerbate exclusion and poverty.

- **Efficiency *versus* safety and sustainability**: pursuing technological progress as quickly as possible may not leave enough time to ensure that developments are safe, robust and reliable.

- Source:[1] *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research*. Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Trustworthy AI

Requirements of Trustworthy AI

source: [2] *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

1 Human agency and oversight

Including fundamental rights, human agency and human oversight

2 Technical robustness and safety

Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility

3 Privacy and data governance

Including respect for privacy, quality and integrity of data, and access to data

4 Transparency

Including traceability, explainability and communication

Requirements of Trustworthy AI

5 Diversity, non-discrimination and fairness

Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation

6 Societal and environmental wellbeing

Including sustainability and environmental friendliness, social impact, society and democracy

7 Accountability

Including auditability, minimisation and reporting of negative impact, trade-offs and redress.


Human agency and Oversight
Technical robustness and Safety
Privacy and Data Governance
Transparency
Diversity, Non-Discrimination and Fairness
Societal and Environmental wellbeing
Accountability
To be continuously evaluated and addressed throughout the AI system's life cycle

Who is responsible?

AI system designers and their managers do have ethical responsibilities.

and

Other stakeholders (e.g. policy makers, politicians, opinion leaders, educators) do have ethical responsibilities.

(Non) Ethical People and (Non) Ethical AI

"I think ethical software development for AI is not fundamentally different from ethical software development in general.

The interesting new question is: ***when AIs learn by themselves, how do we keep them from going astray?***

Fixed rules of ethics, like Asimov's three laws of robotics, are too rigid and fail easily. (That's what his robot stories were about.)

--Pedro Domingos (*Professor at University of Washington*)

- (*) Source: **On Artificial Intelligence, Machine Learning, and Deep Learning. Interview with Pedro Domingos**, ODBMS Industry Watch, June 18, 2018

What do we mean with Ethics?

"So maybe AI will force us to confront what we really mean by ethics before we can decide how we want AIs to be ethical."

-- Pedro Domingos (*Professor at University of Washington*)

- (*) Source: **On Artificial Intelligence, Machine Learning, and Deep Learning. Interview with Pedro Domingos**, ODBMS Industry Watch, June 18, 2018

Some more info …

Course Description

The course will offer a series of lectures covering topics related to Ethics and AI:.

Ethics, Moral Values, Humankind
Trust
Fairness/bias/discrimination;
Transparencies / Explainability/ intelligibility/interpretability;
Privacy/ Responsibility/Accountability, Safety;
Human-in the loop;
Legal Aspects;
Ethics AI in healthcare and other domains.

Course Schedule

- 22.04.2020 **The Ethics of Artificial Intelligence** (AI) (Prof. Roberto V. Zicari)
- 23.04.2020 **The Ethics of Artificial Intelli**gence (AI) (Dr. Emmanuel Goffi)
- 29.04.2020 **Ethics, Moral Values, Humankind, Technology, AI Examples.** (Prof. Rafael A. Calvo)
- **30.04.2020 Ethics, Moral Values, Humankind, Technology, AI** Examples. (Dr. Emmanuel Goffi)
- 06.05.2020 **On the ethics of algorithmic decision-making in healthcare** (Dr. Thomas Grote)
- 07.05.2020 **Fairness, Bias and Discrimination in AI** (Prof. Gemma Roig)
- 13.05.2020 **AI and Trust: Explainability, Transparency** (Prof. Dragutin Petkovic)

Course Schedule

- 14.05.2020 **AI Privacy/ Responsibility/Accountability; Safety; Human-in the loop** (Dr. Magnus Westerlund)
- 20.05.2020 **AI and Trust** (Dr. Christopher Burr)
- 27.05.2020 **Introduction to Z-inspection. A framework to assess Ethical AI** (Prof. Roberto V. Zicari)
- 28.05.2020 **Legal relevance of AI Ethics** (Prof. Florian Möslein)
- 03.06.2020 **AI Fairness and AI Explainability software tools** (Romeo Kienzler)
- 04.06.2020 **Design of Ethics Tools for AI Developers** (Carl-Maria Mörch)
- 10.06.2020 **Assessing AI use cases. Ethical tensions, Trade offs.**
- 17.06.2020 **Assessing AI use cases. Ethical tensions, Trade offs**.
- 18.06.2020 **Assessing AI use cases. Ethical tensions, Trade offs.**(Dr. Estella Hebert)

Admin

Remote, starts on **April 22 at 2pm SHARP (Berlin time- CET) via Zoom.**

For all people who registered, prior to each lecture, you will receive by e-mail a link to join the Zoom video call.

Time: Wednesday and Thursday from 2pm (SHARP!) to 4pm (Berlin time- CET)

Language: The language of the lectures is **English**

Credit Points: Students can receive **6 CPs**. Link in QIS/LFS

Module names: DC, M-DS-ADS, B-WB, M-WB, PoE, M-SIW-I1A, M-SIW-I1B

Eligibility: Bachelor Students, Master Students, and PhD students across multiple disciplines are encouraged to attend.

– Open to all interested people. – No fee required.

Communication via Email: EthicalAISS2020@gmail.com

Credit Points and Assignments.

- **Registration: Deadline April 20, 2020.**
- **For each registered student, we create a Google Doc** (link sent by e-mail)
- **Assignments**:

Each student, individually, will read 1 selected report/paper every week. For each paper you will need to answer two questions in written form.

For each question satisfactorily answered you get 1 Point.

In order to get the final credit points you need to have at least a total of 7 Points.

Credit Points and Assignments.

- Updates per week: 1 paper to read/ 2 questions to answer.
- Deadline to produce the answer: 1 week (on next Wednesday at 10am).
- Failure to submit on time, no points for the Qs.
- 1 week for us to correct answers (results into student Google doc)

Credit Points and Assignments.

Answer to Questions :

1. In written form. English only.

2. Min. ½ page text, Max. 1 page text.

3. Optional: add max 1/2 page (figures).

4. Always quote the source! No copy and paste (if discovered copy and paste, with no source, no CPs).

Grades (for Students Goethe University)

7 papers (7 weeks),
2 Qs per paper: 14 Qs
1 Point per Q: max 14 points

Grades:	Points
1	13 - 14 points
2	11 - 12 points
3	9 - 10 points
4	7 - 8 points
5	Less than 7 points

On the road…