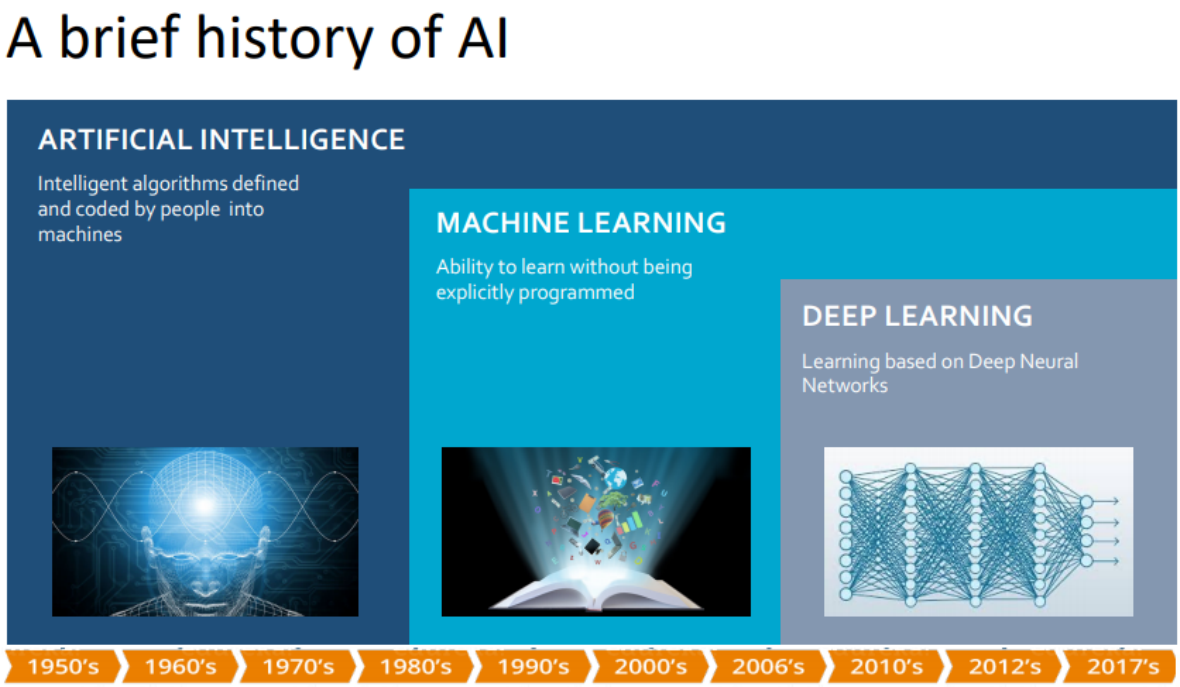
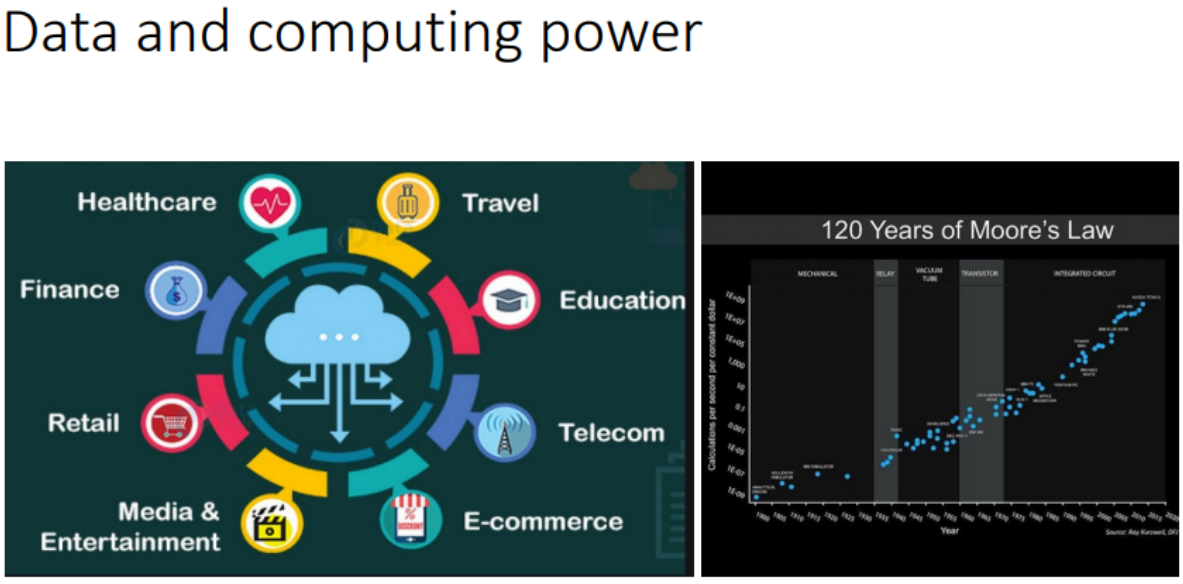
Lesson_14_IBM's_swag



Many of the AI ethics concerns are about the misuse of the technology, but some of them are related to specific techiniques.



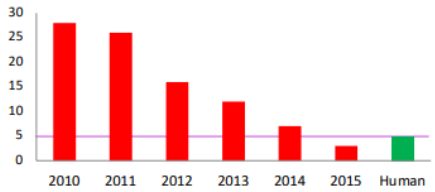
The increased use of Machine learning techniques, which are the ones most related to AI ethics concerns, is due to the presence of a huge amount of data in different

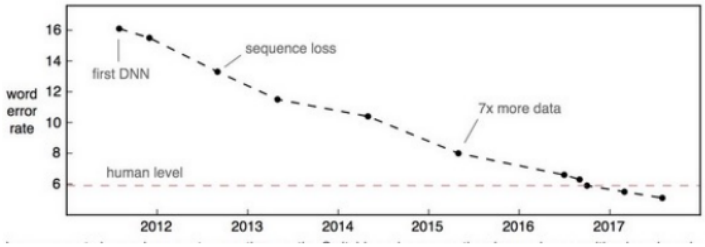
areas and also the computing power to deal with that amount.

Image and natural language interpretation

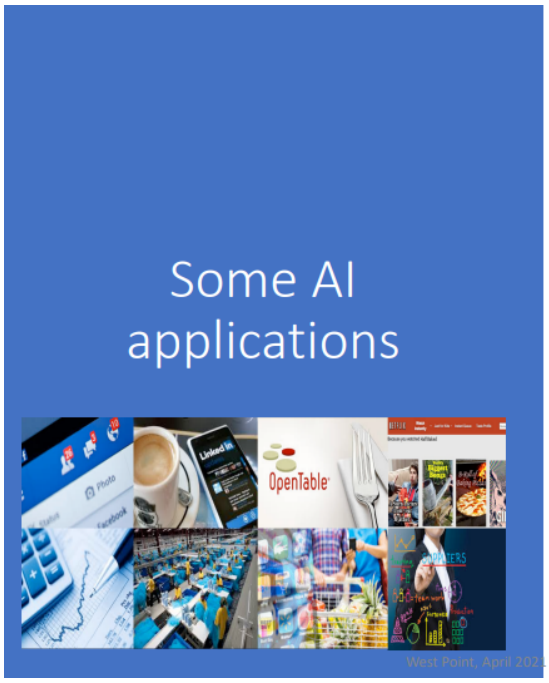
Woman holding a cask of bananas

A group of young people playing fresbee





This led to a lot of success in many application where the AI system perceives the external world and then responds to it (e.g, image interpretation, speech recognition ...). Already few years ago machines were able to outperform humans in a specific task.



- Digital assistants:
  - Home assistants (Alexa)
  - Travel assistants (Waze)
- Driving/travel support:
  - Auto-pilot (Tesla)
  - Ride-sharing apps (Uber, Lyft)
- Customer care:
  - Client service chatbots
- Online recommendations:
  - Friend recommendations (Facebook)
  - Purchase recommendations (Amazon)
  - Movie recommendations (Netflix)
- Media and news:
  - Ad placement (Google)
  - News curation
- Healthcare:
  - Medical image analysis
  - Treatment plan recommendation
- Financial services:
  - Credit risk scoring
  - Loan approval
  - Fraud detection
- Job market:
  - Resume prioritization
- Judicial system:
  - Recidivism prediction (Compas)

IBM

Many applications derive from the machines' capabilities.

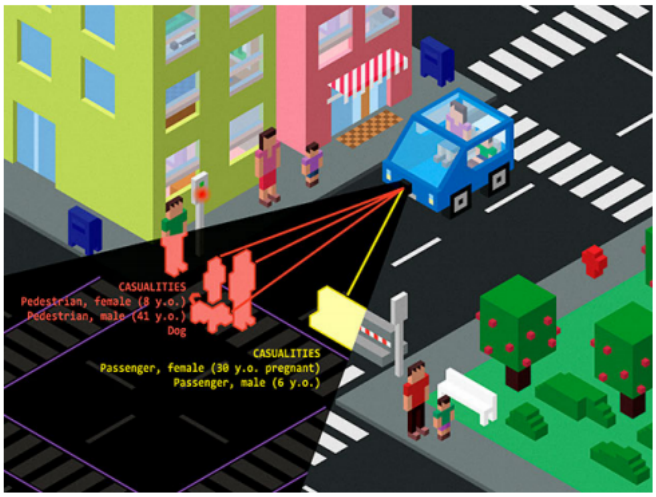
AI has a lot of capabilities, but also a lot of limitations. The lack of robustness and adaptability are related: AI focused on a specific problem lacks the capability of transfering the learned knowledge to another problem.

The errors in image classification can be due to noise or other external interventions, which would not cause trouble to humans, are another sign of the lack of robustness. These are the reasons why there is reluctancy of using these models where some danger may derive from these errors or there's need of accountability.

Ethical issues -- examples

Ethical issues related to AI in some companies

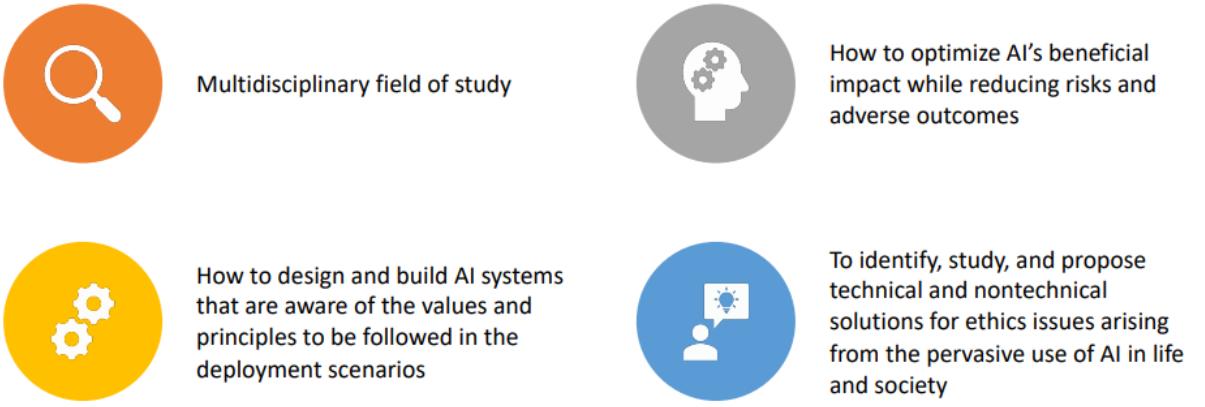
Can we trust AI's decisions?



Usual Trolley Problem with self-driving cars: harm pedestrians or the passengers?

On what values should be the car programmed on for us to be able to trust the decisions it will make?

AI Ethics

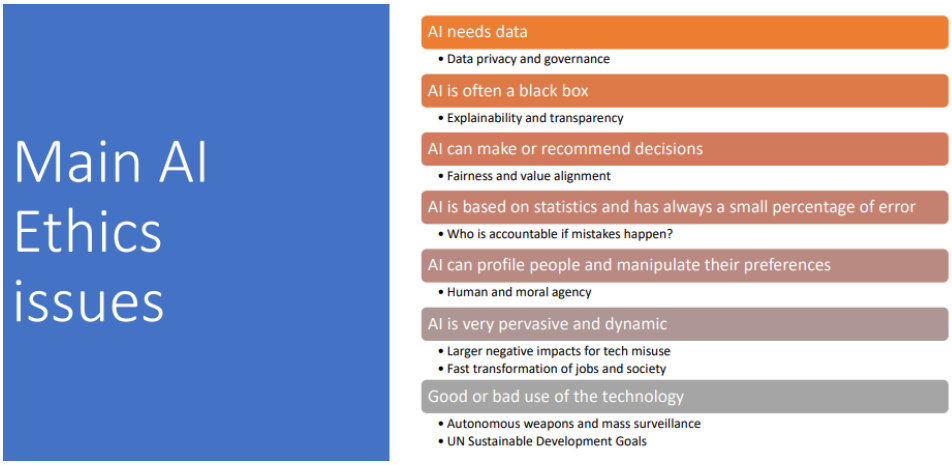


AI ethics is a multidisciplinary field of study that involves not only AI experts but other experts who are able to assess the **impact** that the new technologies can have on society (sociologist, philosophers, economist...). Also policy makers, who choose the regulations of our every-day life.

The goals of this field of study are:

- **Optimize** AI's beneficial impact, while reducing risks.

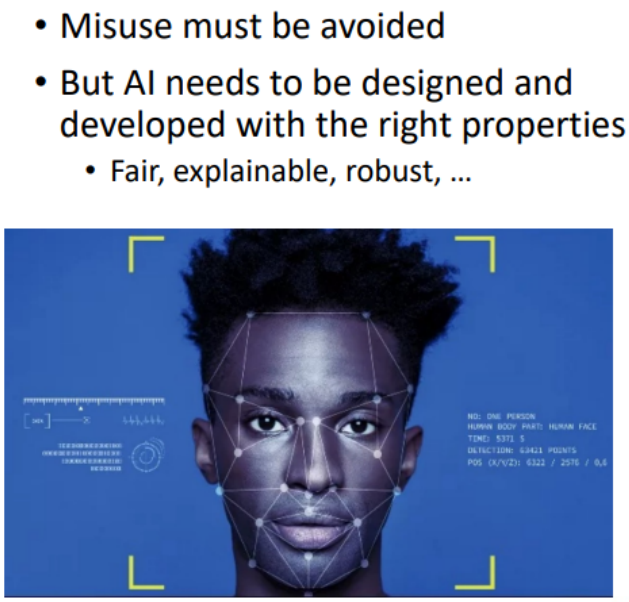
- Build machines that will work according to some values **relevant** for their deployment scenario.
- Development of guidelines, best-practices, standards, regulations and laws; these can help improving the technical solutions and, maybe, completing them.



AI needs raise ethical problems:

- **Data privacy & Governance**: the gathering of data involves personal and sensitive data, what happens to them? Who gets them?
- **Explainability & Transparency:** how did the machine came up with a certain conclusion is hardly explainable. Also, they're not very transparent: how they been built? What design and princples have been followed?
  They are kind of a black box and this generates concerns and mistrust.
- **Fairness & alignement:** the AI systems can make recomendations, but how can we be sure that they are aware of human values or aligned to them? We want the decisions to be **fair** and not create discrimination in the same group of people, for example.
- **Accountability:** we know that the machine learning systems are not perfect and will commit some errors, like humans do. But when a machine makes a mistake who's to be held accountable?
- **Profiling & Manipulation:** profiling is done to infer an individual's preferences, in order to show advertisment that we are interested in. If our preferences are variegated, it's hard to understand what we may like or not. But if they are polarized, then is easier. Providing us content that may steer our preferences towards polarization is a form of manipulation.

- **AI pervasiveness:** the high pervasiveness of AI implies that a misuse of the technology has a very large impact across different application domains, and also brings many transformations (jobs, society ...). Are we keeping track of these transformations, are the possibly negative aspects interpreted so that we can put in place some solutions?

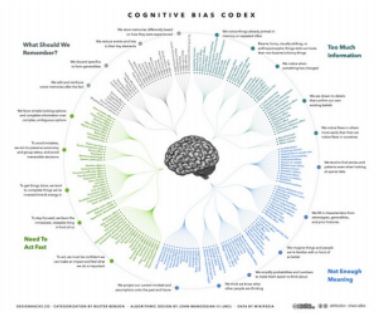


Misuse must be avoided of course, but AI technologies must be developed with the right properties in order to be sure that the way we deal with AI is ethical and responsible.

AI is not a neutral technology: the way it is designed and developed and the properties that are embedded into it are also part of how the technology will behave, It's not just a matter of avoiding misuse.

AI fairness

- Bias: prejudice for or against something
- As a consequence of bias, one could behave unfairly to certain groups compared to others
- Why should AI be biased?
  - Trained on data provided by people, and people are biased



Very similar to the logic in the Fairness in algorithmic Decision-making

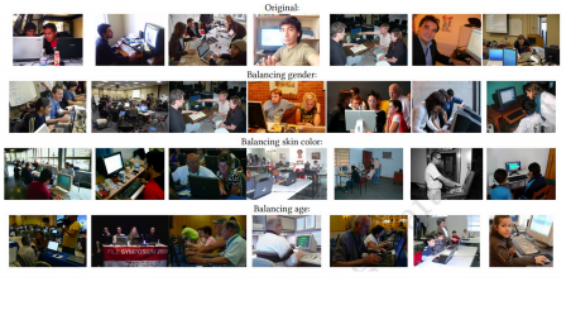
Why should be AI biased? Biases lead to discriminatory behaviour

People who build the AI will also generate the data on which it trains. But people are biased... Any step involving humans during the development will introduce some biases.

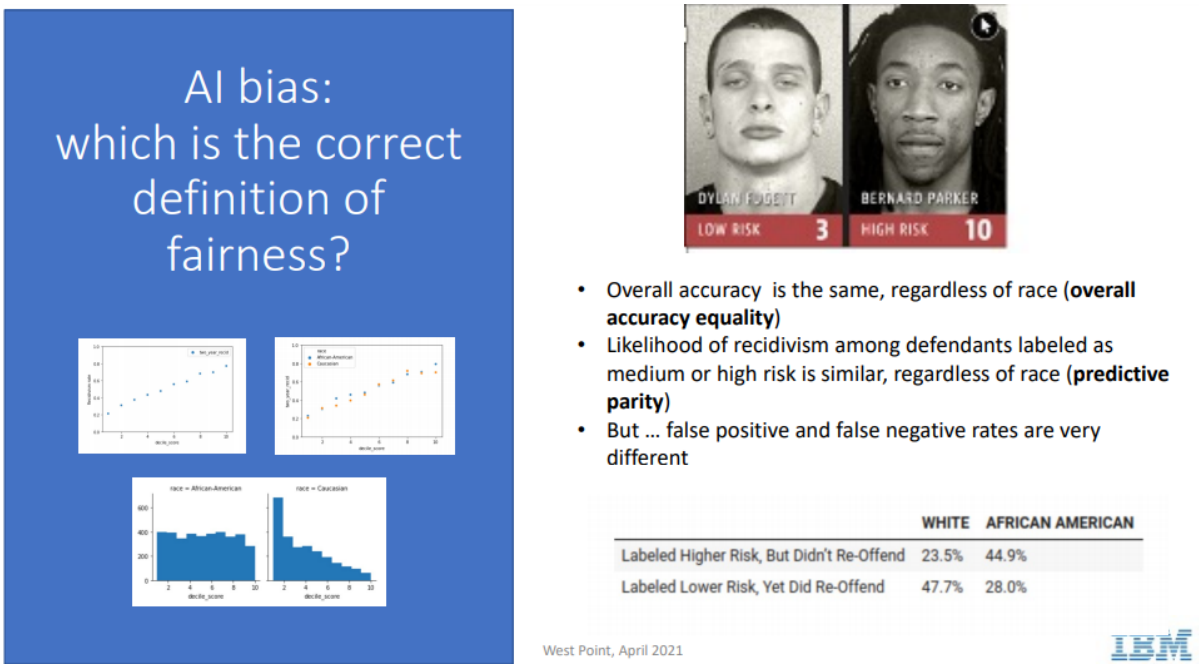
AI bias: ImageNet

- 14M images, used to train image interpretation AI systems
- Bias in the data distribution and in the data labels (Mturk people)

IMAGENET



Biases come from not only from the data but also from every decision we make. For example, if we build a mortgage application acceptance AI system: given a mortgage application it decides wheter it should be accepted or declined. During the development of the system, we might want to introduce some prioritized motivations for loan applications. Given this priorities, we assume to know what is best for most people. But what might be best for one ethnicity might not be for another one. Therefore, these priorities introduce a form of bias.

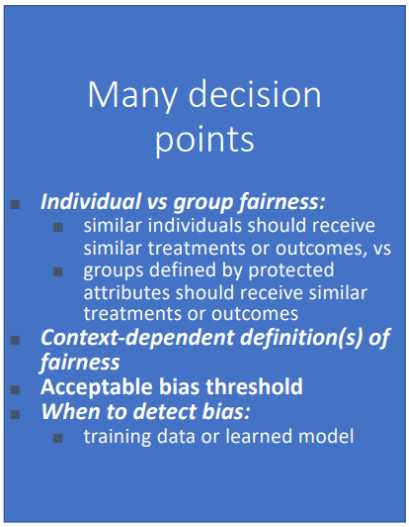


COMPAS System again

The algorithm's output is a score that describes the probability that the subject will commit a crime again. The judges would use these scores to decide if the subject should be staying in jail while waiting for the final judgement.

The algorithm has been found to be biased, but it highlighted the fact that there are more definition of fairness: an algorithm can be considered fair according to one notion but not to another one. In this case, the accuracy of the algorithm over the two groups (black & white) was more or less the same: there was equality in overall accuracy. According to this definition of fairness, the algorithm was ok. But looking at the errors, the false positives (high risk when it should be low) and false negative (low when high), it was found that the rates were very different, when they should have been similar at least. According to this definition of fairness, the algorithm was not fair.

This shows that to be able to assess an AI system fairness, we must choose the notion **relevant** for that application domain.



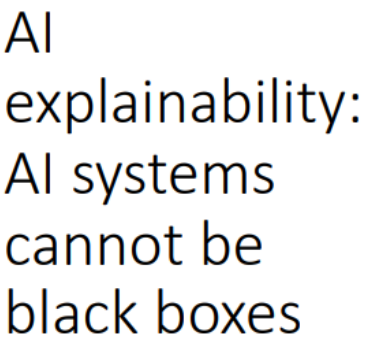
Source: Fairness and Machine Learning by Solon Barocas, Moritz Hardt, Arvind Narayanan (https://www.fairmlbook.org)

Name	Closest relative	Note	Reference
Statistical parity	Independence	Equivalent	Dwork et al. (2011)
Group fairness	Independence	Equivalent	
Demographic parity	Independence	Equivalent	
Conditional statistical parity	Independence	Relaxation	Corbett-Davies et al. (2017)
Darlington criterion (4)	Independence	Equivalent	Darlington (1971)
Equal opportunity	Separation	Relaxation	Hardt, Price, Srebro (2016)
Equalized odds	Separation	Equivalent	Hardt, Price, Srebro (2016)
Conditional procedure accuracy	Separation	Equivalent	Berk et al. (2017)
Avoiding disparate mistreatment	Separation	Equivalent	Zafar et al. (2017)
Balance for the negative class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Balance for the positive class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Predictive equality	Separation	Relaxation	Chouldechova (2016)
Equalized correlations	Separation	Relaxation	Woodworth (2017)
Darlington criterion (3)	Separation	Relaxation	Darlington (1971)
Cleary model	Sufficiency	Equivalent	Cleary (1966)
Conditional use accuracy	Sufficiency	Equivalent	Berk et al. (2017)
Predictive parity	Sufficiency	Relaxation	Chouldechova (2016)
Calibration within groups	Sufficiency	Equivalent	Chouldechova (2016)
Darlington criterion (1), (2)	Sufficiency	Relaxation	Darlington (1971)

West Point, April 2021 IBM

Once we have chosen the definition of fairness, we must find an acceptable threshold for bias since it's unavoidable (80% is the typical one).

The developers should be educated to recognize their own biases and understand what biases they can inject into the system. Another method to keep biases in check is to form heterogenous developing teams, able to spot each others biases.



The data subjects that give the data to the AI system have the **right** to have an explanation of the decisions made by the system.

Again, a system that does not give any explanation for its outcomes it's unlikely to be trusted, in general.

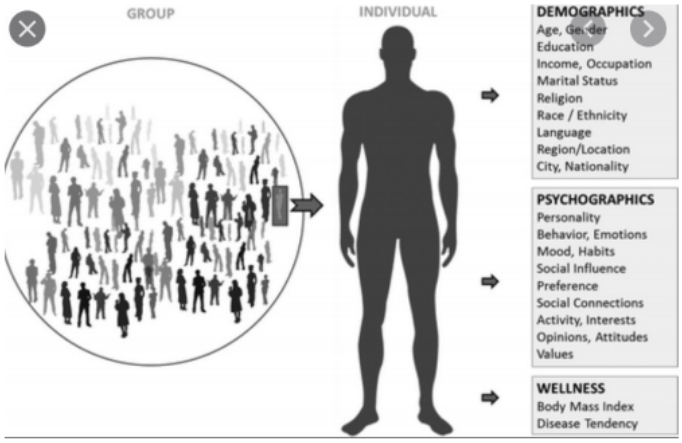
A system that works in a certain domain should be able to generate different explanation for different audiencies. For example, a system working in the medical domain should be able to generate explanation for the doctors, the patients and so on, in a way that is understandable for them.

Data handling: the General Data Protection Regulation (GDPR)

The GDPR is about data protection and sets a standard and gives regulation, very important for EU. Many countries lack a comprehensive document like this one, even if it is partially used in some other regions of the world.

Profiling and manipulation

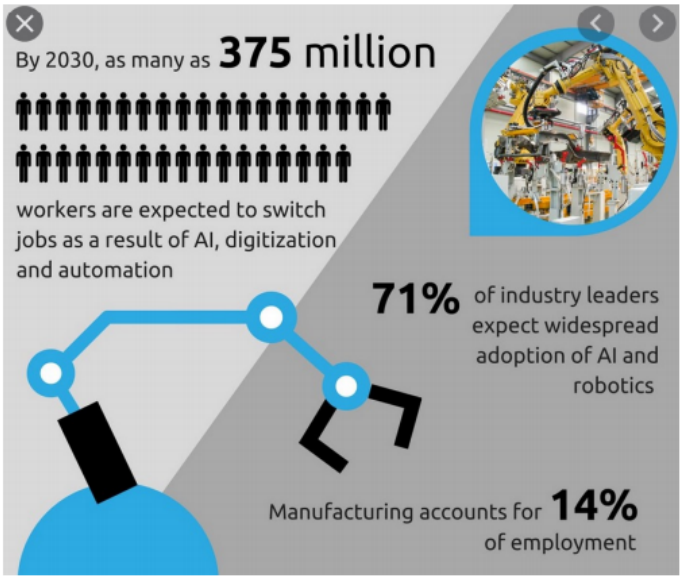
- From actions to profiles
  - Like, text, images, follow, ...
- AI can infer our preferences, and use them to advertise products that we probably like
  - Easier if our preferences are bipolar



As we said before, our preferences are tried to be made more polarized by the advertising systems because it eases the job of giving highly preferred ads.

Impact on the workforce

- Many jobs will disappear, and many others will be created
- All jobs will change

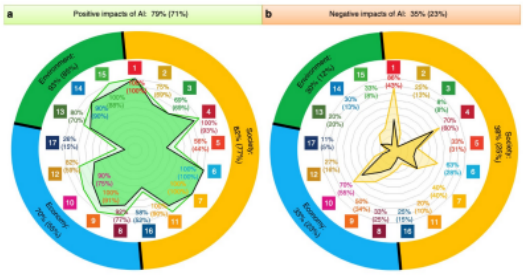


People have to be convinced of the transition benefits because it's easier to imagine existing jobs disappear rather than imagine those which will be created.

It is expected that AI will change the jobs, for now it will be a support for the human professional figures but also will guide robotics in the manufactoring environment.

A vision of the future (2030)

- 17 goals, 169 targets
- Very difficult path
  - The pandemic has worsened the situation
- AI can help in achiving the SDGs
- COVID: vaccines in less than one year!



The UN have put together a comprehensive vision of what it means to improve the world. They used 17 goals that define how the world should look like in 2030. A recent study showed how much the current AI techniques have been used to move towards these SDG and how much it has brought further away from them.

This SDG can help to steer the use of technology in a desirable direction.

Where was IBM when the AI fell?

IBM, technology, and AI

- 110 years
- Hardware e software
- Enterprise AI: AI solutions for other companies
  - Banks and financial institutions
  - Governments
  - Aeroports
  - Hospitals
  - ...

Summit, IBM

Chess: IBM Deep Blue, 1997

Quantum computer, IBM

Jeopardy: IBM Watson, 2011

Project Debater, 2020

Overview of the IBM technology in the AI field, interesting but doesn't seem to be strictly related to ethics or whaterever else relevant for the course.

IBM Principles of Trust and Transparency (2017)

The purpose of AI is to augment human intelligence

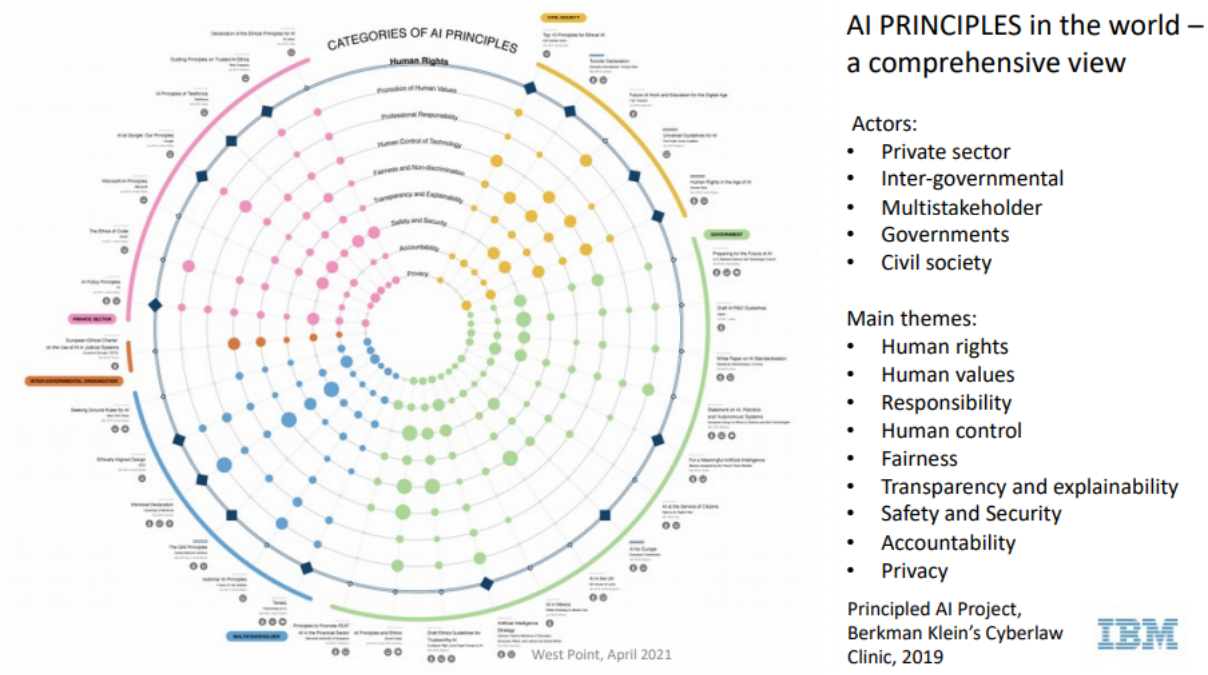
Data and insights belong to their creator

New technology, including AI systems, must be transparent and explainable

IBM in terms of ethics defined in 2017 some high-level principles:

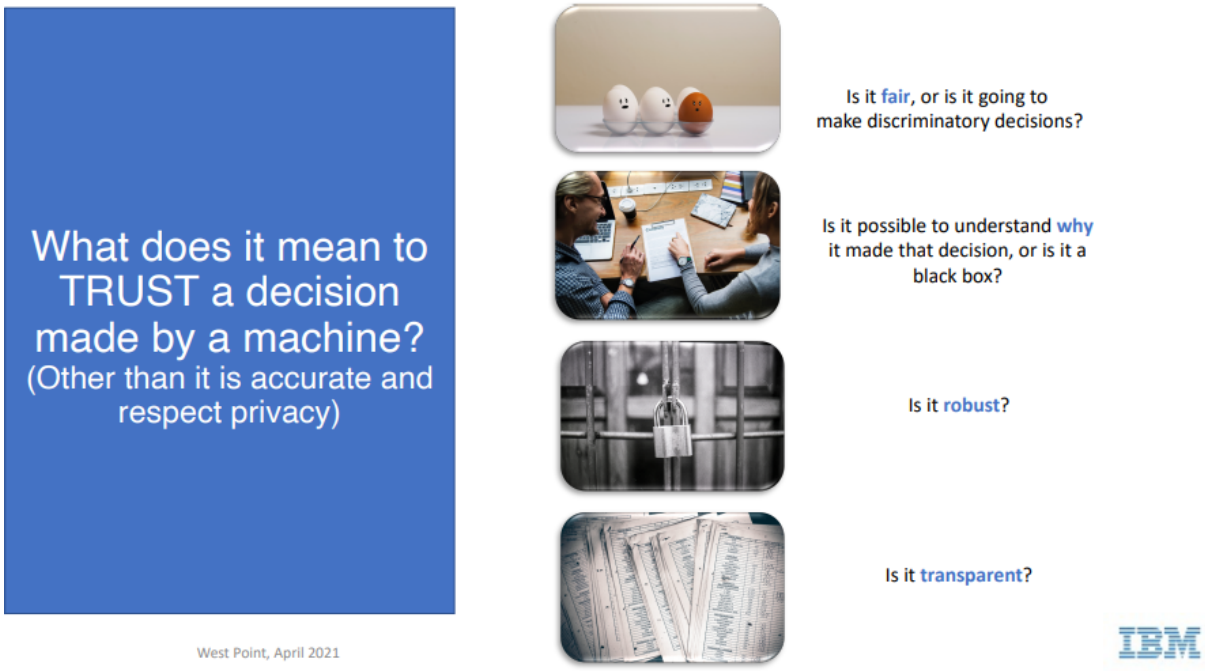
- AI should a*ugment* human intellegence rather than replace it, therefore helping humans in their decisions.

- Data gathered from clients belong to them: they will not be reused to improve solutions for other clients.
- The technology should be transparent, for example a human being should always know whether s/he is interacting with a human or an AI system. The technology should be explainable

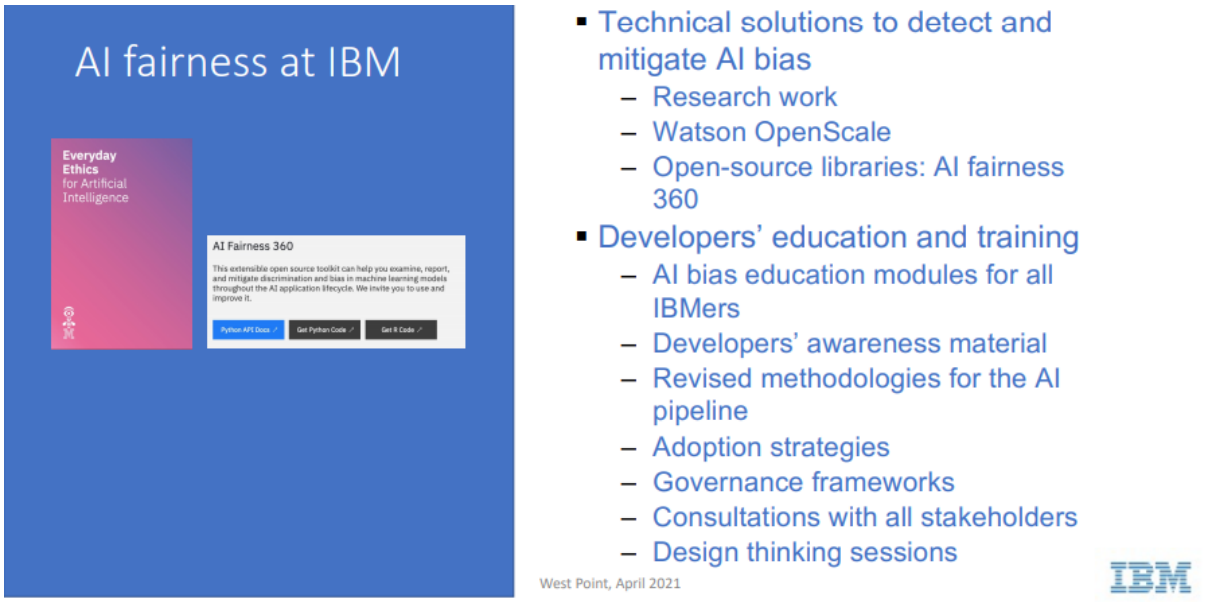


Each ray represents one set of principles while the colors represent actors

These principles were the starting point of a long path and from the principles now we should take some concrete actions, implementing them in the technology.



Trust definition using these four pillars listed above



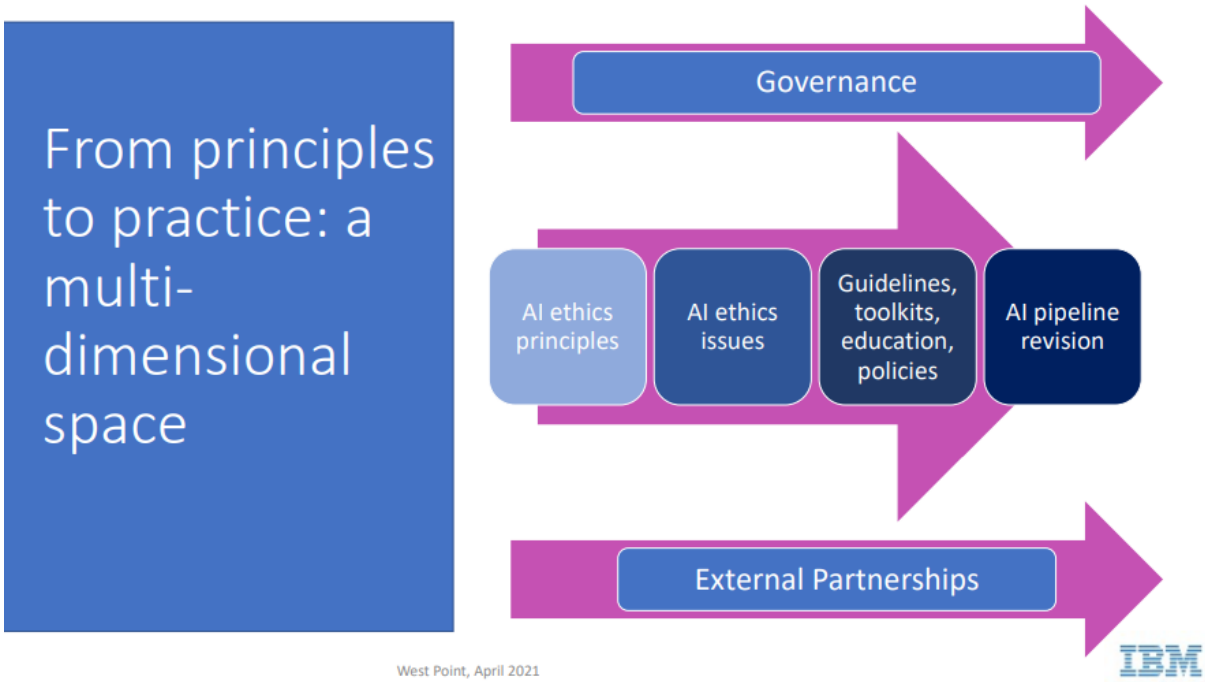
IBM's initiatives to respect the defined principles

Other than tools to inspect and detect unfair behaviours, IBM investend into educating its developers to help them undestand those tools, when to used them and why.

- AI factsheet
  - Transparency by documentation
  - Design a development choices
  - Not just a checklist
  - Self-assessment and beyond
- Useful to
  - Developers
  - Clients
  - Users regulators/auditors
- Aligned with EC High Level Expert Group on AI self-assessment list (ALTAI)
- AI factsheet 360

IBM wants to be transparent on all the design decisions that it has made during the development process and to do that uses AI factsheet, which are an implementation that follows the idea of ***transparency by documentation***.

AI factsheet 360 open-source toolkit to understand this concept and possibly use them.



It is a journey, it takes a long time to change how things are done, especially in big companies.

Governance: the IBM AI Ethics board

- Mission
  - Awareness and coordination
  - Internal education and retraining
  - Linking research to services and platforms
  - Advice to business units
  - Internal governance framework
  - Define policies and advice regulators
- Risk-based approach for the BUs
  - Vetting based on three dimensions (tech, use, client)

This board supervises all the mentioned activities and it decides whether if an AI solution can be offered to a certain client, if there are ethical uncertainties about it. The board takes that decision according to certain parameters (e.g, Bias threshold), considering what the deployment the solution will have and its context and the client itself.

Partnerships

Academia
Companies
Governments
Civil society organizations

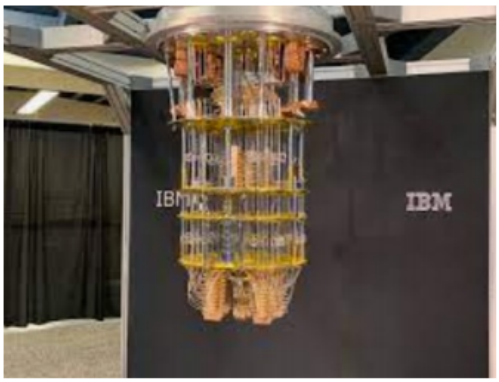
Multi-disciplinary and multi-stakeholder

West Point, April 2021

Partnerships with external entity is necessary to better understand the issues: it has to be multi-disciplinary, but also multi-stakeholder because all the different voices have to be there to understand the impact on society.

Not just AI

- Neurotechnologies
  - Huge potential for healthcare
  - Reading/writing neurodata
  - Additional issues around privacy, agency, and identity
- Quantum computing
  - How to responsibly use such a huge computing power?



There are other fields that can combined with AI raise ethical concerns: one example is **neurotechnologies**; what would happen to the concept of privacy if we were able to read neurodata (nervous system's signals)?

Quantum computing will provide huge computational power and some protection based on complexity may fall: how can we protect it from cyber attacks? How can we ensure that this technology will be used in a responsible way?

List of useful links in the slides.