AI Ethics at IBM:
From Principles to Practice

Francesca Rossi

IBM fellow and AI Ethics Global Leader

MAI4CAREU | Master programmes in Artificial Intelligence 4 Careers in Europe

IBM

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

A brief history of AI

ARTIFICIAL INTELLIGENCE

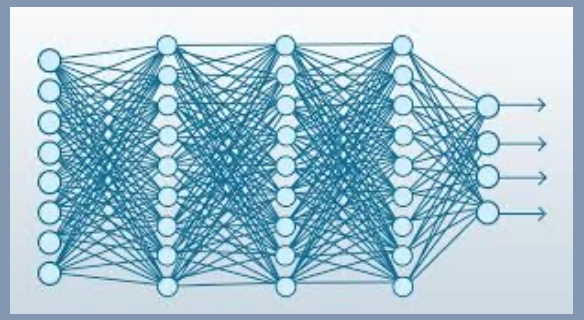
Intelligent algorithms defined and coded by people into machines

MACHINE LEARNING

Ability to learn without being explicitly programmed

DEEP LEARNING

Learning based on Deep Neural Networks



1950's 1960's 1970's 1980's 1990's 2000's 2006's 2010's 2012's 2017's

IBM

Data and computing power

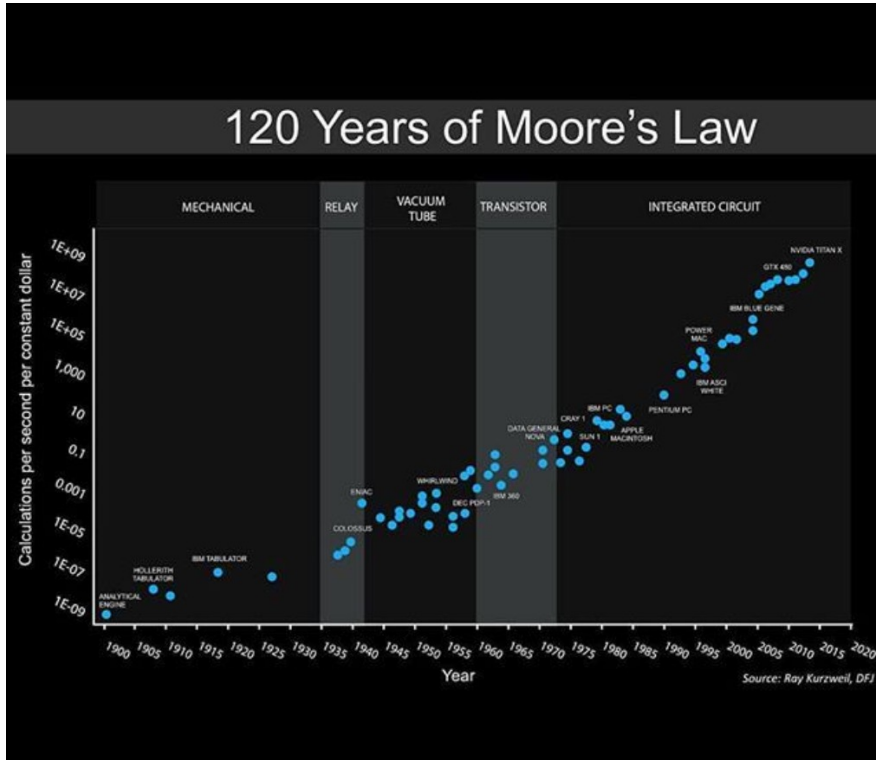
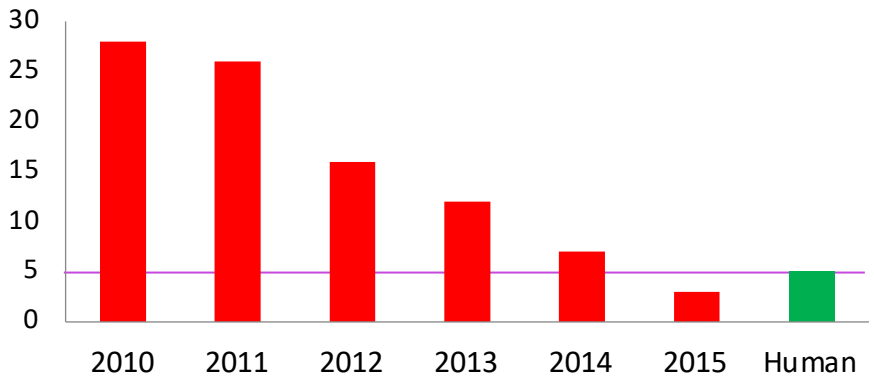
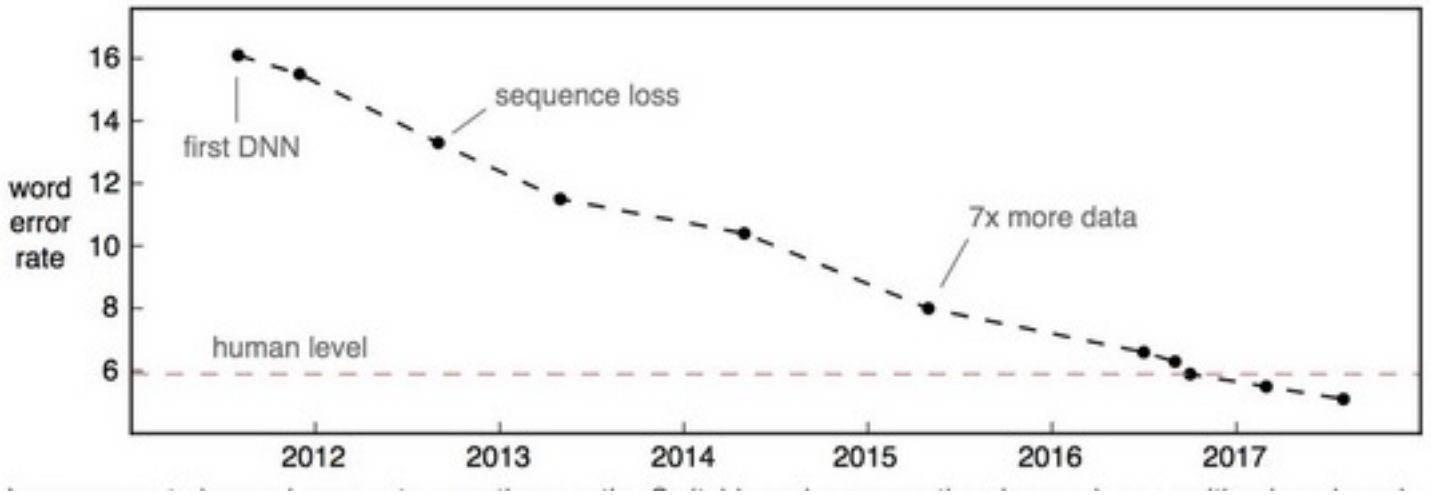


Image and natural language interpretation

Woman holding a cask of bananas

A group of young people playing fresbee

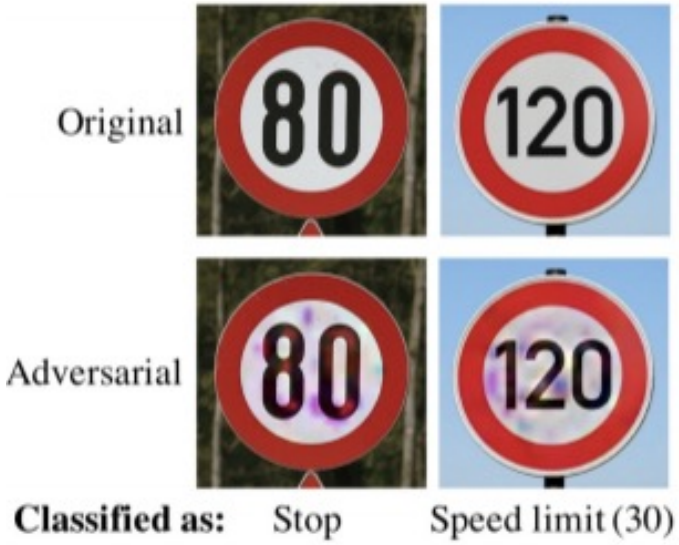


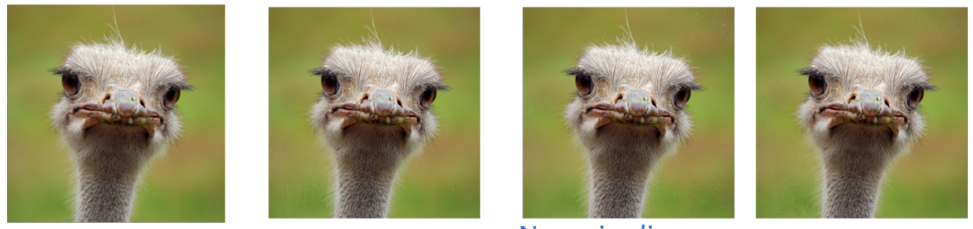


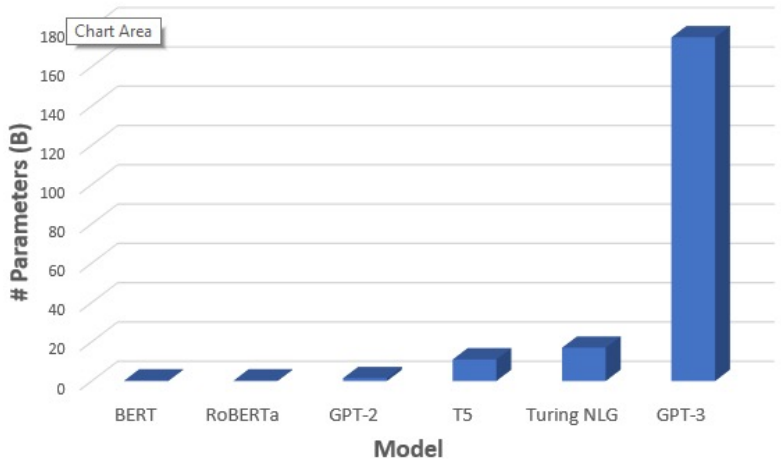
Some AI applications

- 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







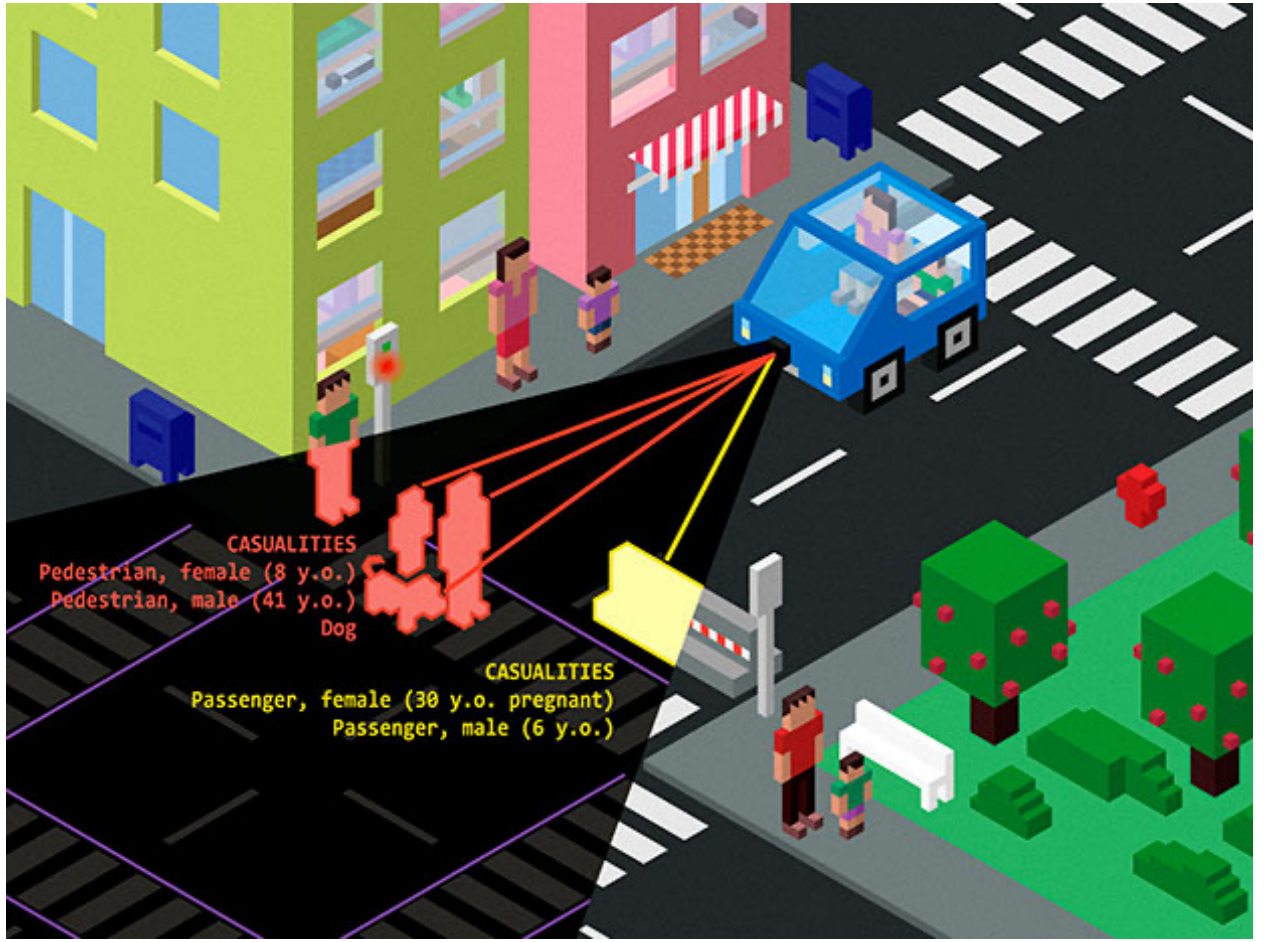
AI limitations

- Narrow AI
  - Solves well specific problems
- Lack of robustness and adaptability
- Needs a lot of resources
  - Data and computing power

Ethical issues -- examples

IBM Confidential

Can we trust AI's decisions?



AI Ethics

Multidisciplinary field of study

How to optimize AI's beneficial impact while reducing risks and adverse outcomes

How to design and build AI systems that are aware of the values and principles to be followed in the deployment scenarios

To identify, study, and propose technical and nontechnical solutions for ethics issues arising from the pervasive use of AI in life and society

IBM

Main AI Ethics issues

AI needs data

- Data privacy and governance

AI is often a black box

- Explainability and transparency

AI can make or recommend decisions

- Fairness and value alignment

AI is based on statistics and has always a small percentage of error

- Who is accountable if mistakes happen?

AI can profile people and manipulate their preferences

- Human and moral agency

AI is very pervasive and dynamic

- Larger negative impacts for tech misuse
- Fast transformation of jobs and society

Good or bad use of the technology

- Autonomous weapons and mass surveillance
- UN Sustainable Development Goals

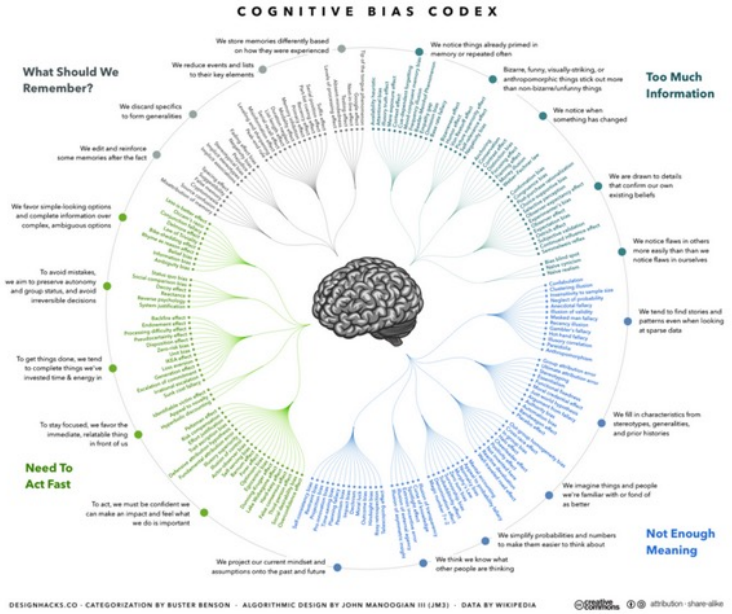
IBM

AI is not a neutral technology

- Misuse must be avoided
- But AI needs to be designed and developed with the right properties
  - Fair, explainable, robust, …

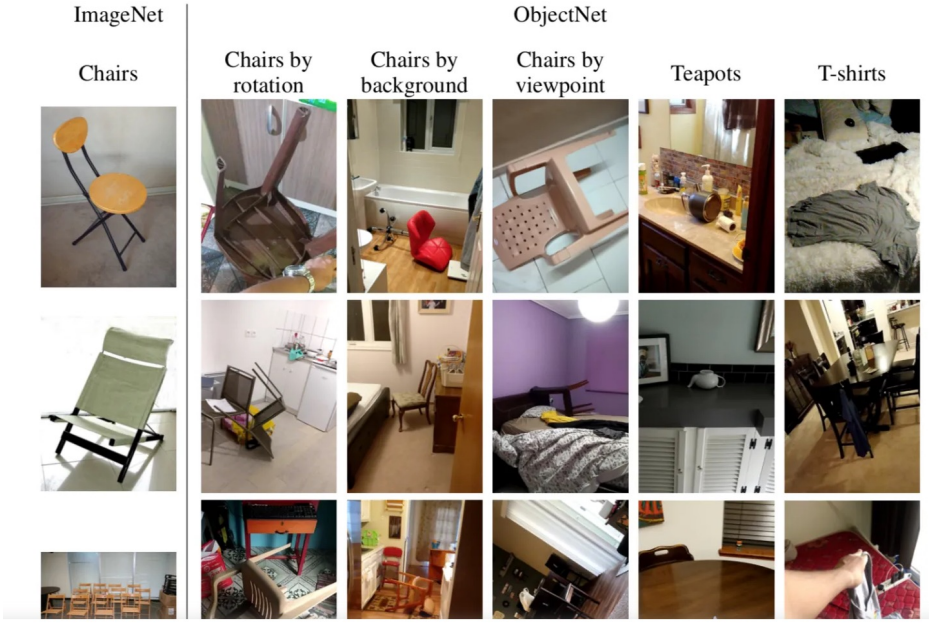
AI fairness

- 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



AI bias: ImageNet

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

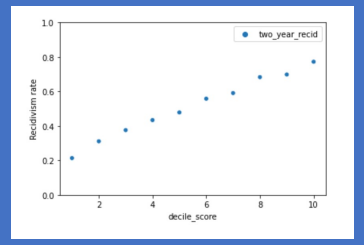


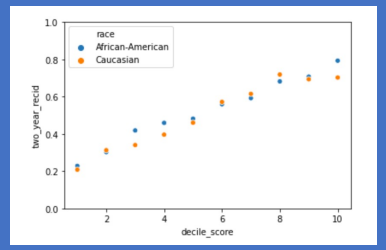
Mortgage application: bias not just from data

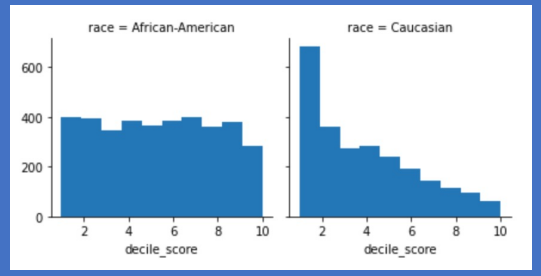
- Training data
  - Ex. : correlation gender-acceptance
- Design decisions:
  - Ex.: prioritized motivations for loan applications
  - Buying a house
  - Paying school fees
  - Paying legal fees
    - Loan applications with these motivations are prioritized
    - If one of them is omitted, the relevant community will be penalized

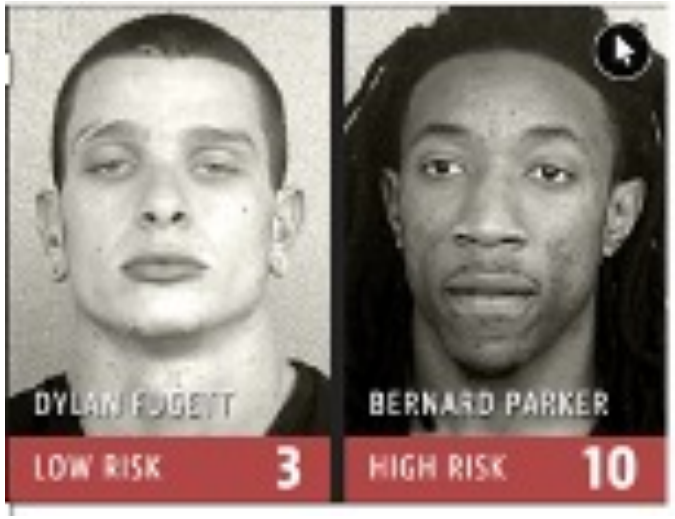
IBM

AI bias: which is the correct definition of fairness?









- Overall accuracy is the same, regardless of race (**overall accuracy equality**)
- Likelihood of recidivism among defendants labeled as medium or high risk is similar, regardless of race (**predictive parity**)
- But … false positive and false negative rates are very different

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

IBM

Many decision points

- ***Individual vs group fairness:***
  - similar individuals should receive similar treatments or outcomes, vs
  - groups defined by protected attributes should receive similar treatments or outcomes
- ***Context-dependent definition(s) of fairness***
- **Acceptable bias threshold**
- ***When to detect bias:***
  - training data or learned model

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)

AI explainability: AI systems cannot be black boxes

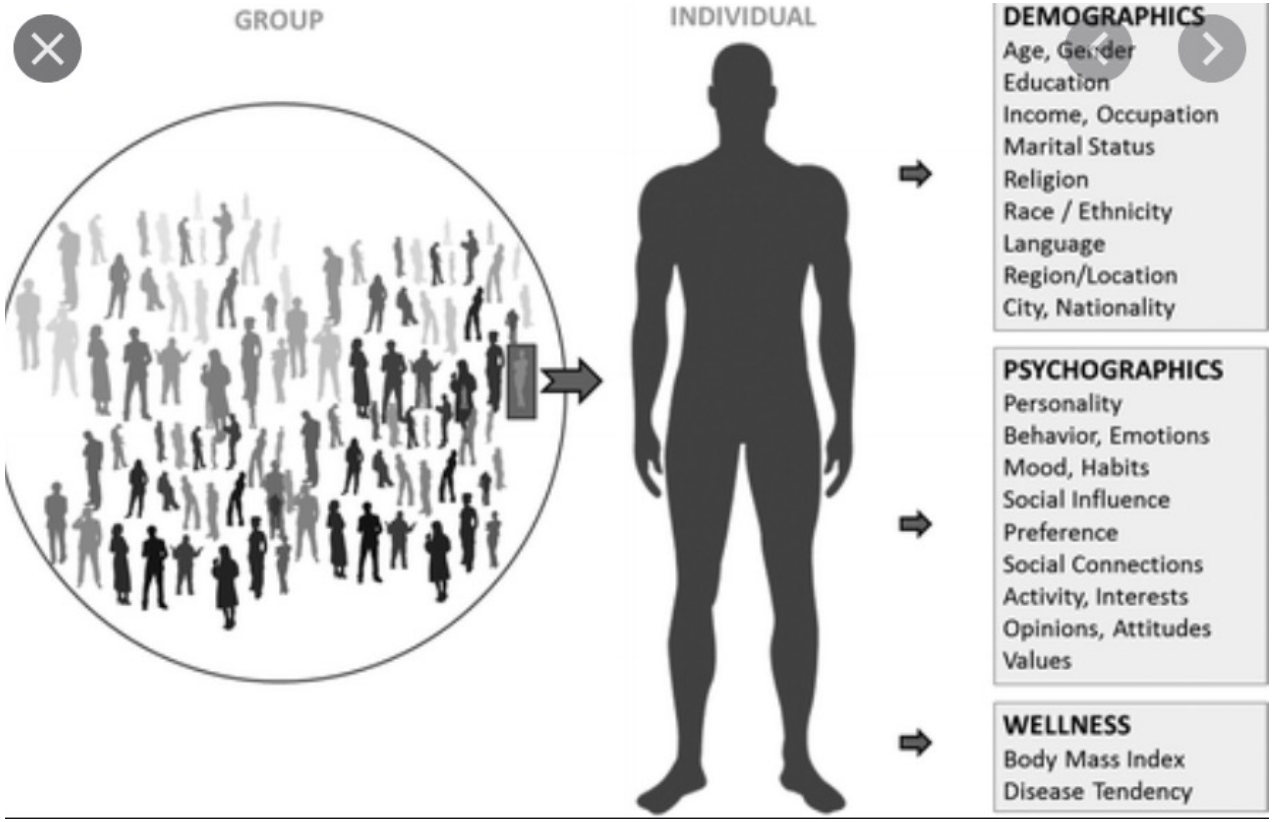
The General Data Protection Regulation (GDPR)

- Limits to decision-making based solely on automated processing and profiling (Art.22)
- Right to be provided with meaningful information about the logic involved in the decision (Art.13 (2) f. and 15 (1) h)

Data handling: the General Data Protection Regulation (GDPR)

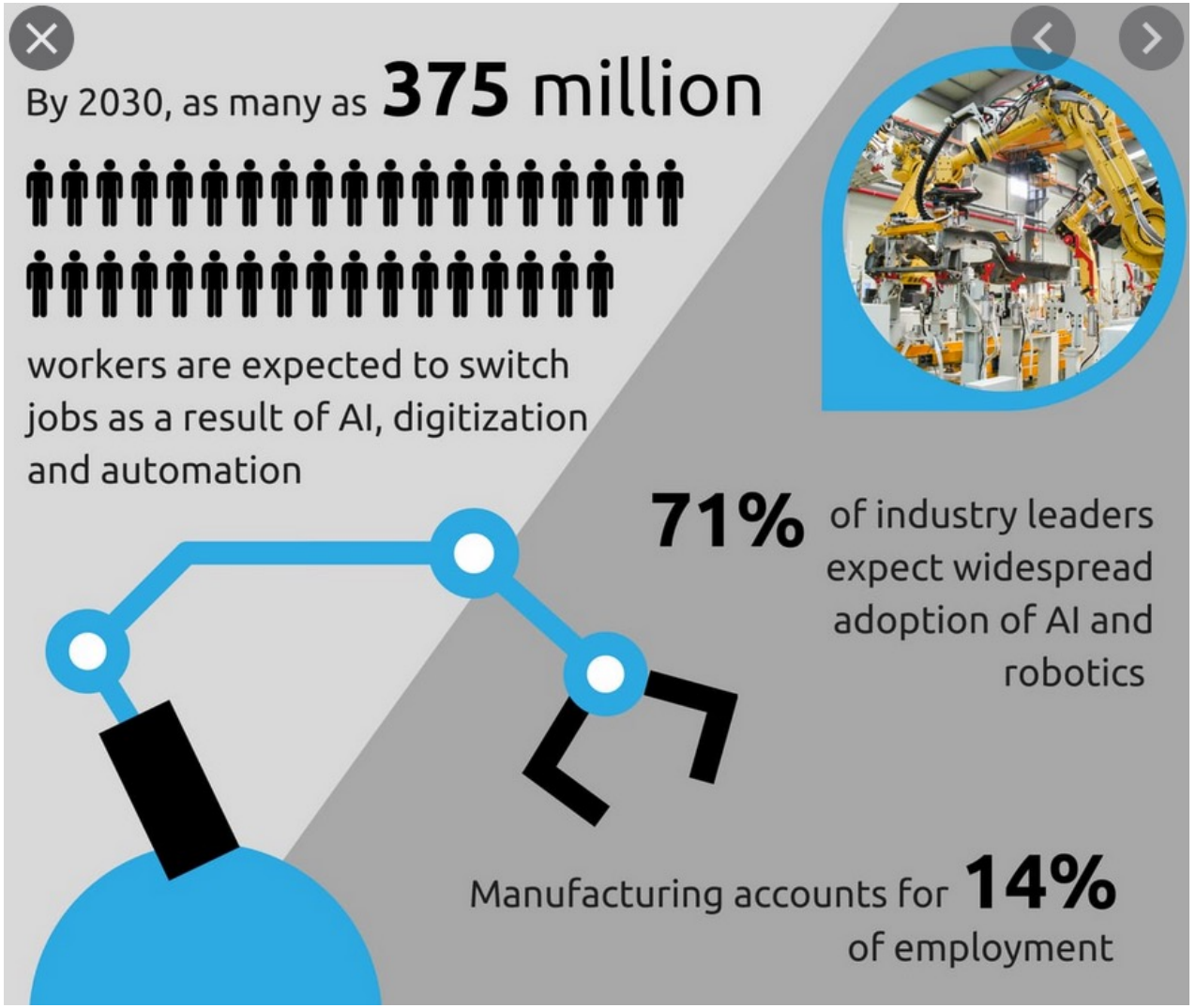
Profiling and manipulation

- 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



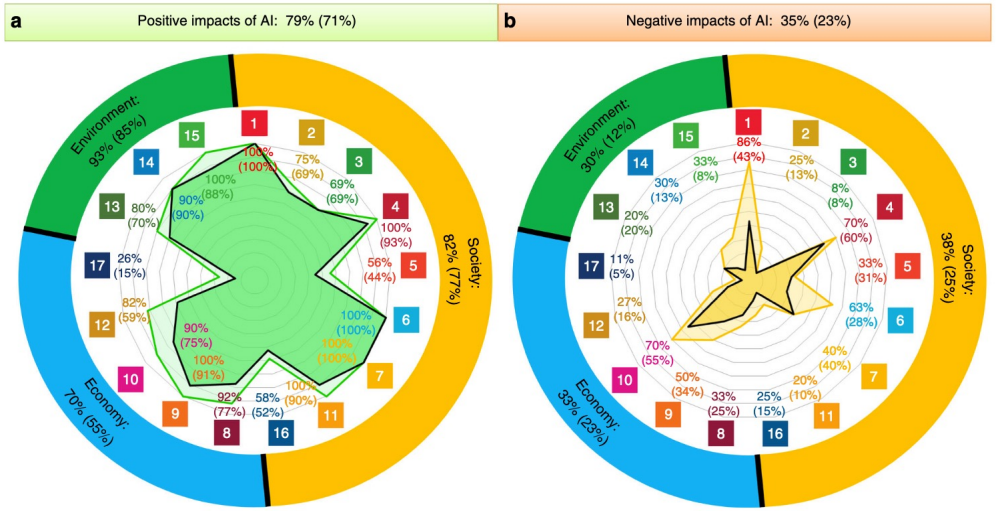
Impact on the workforce

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



A vision of the future (2030)

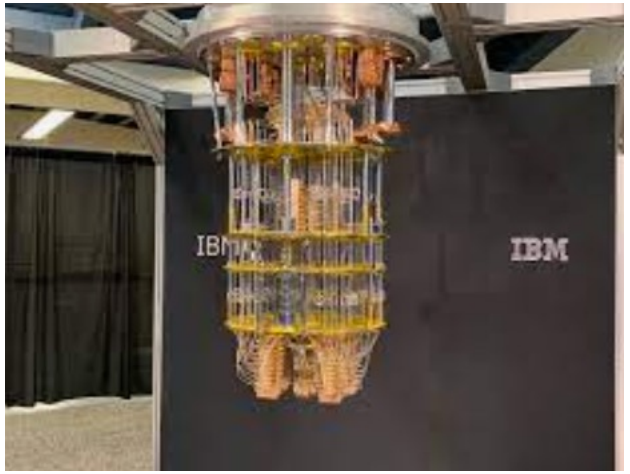
- 17 goals, 169 targets
- Very difficult path
  - The pandemic has worsened the situation
- AI can help in achiving the SDGs
- COVID: vaccines in less than one year!



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



Quantum computer, IBM

Chess: IBM Deep Blue, 1997

Jeopardy: IBM Watson, 2011

Project Debater, 2020

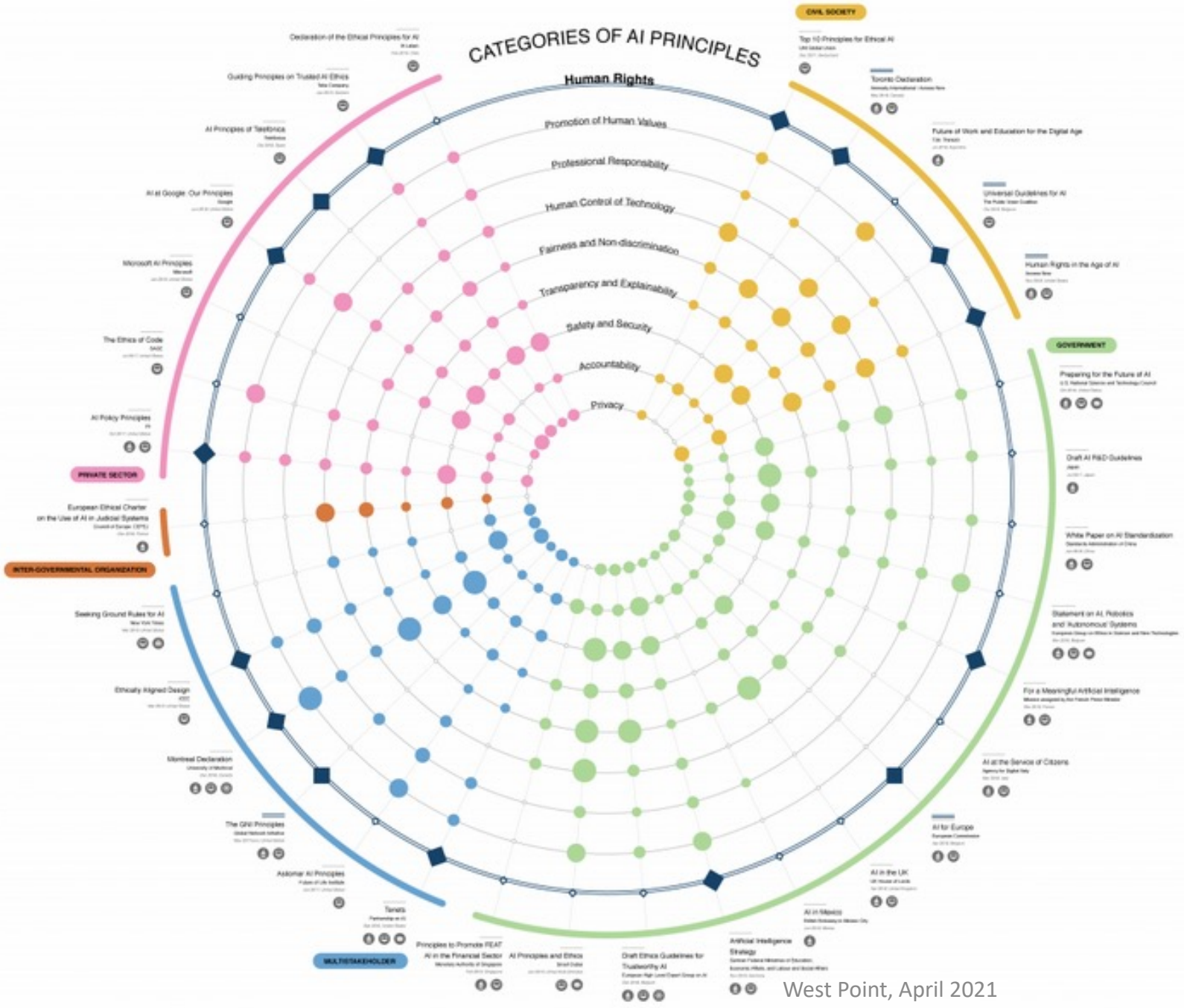
IBM Principles of Trust and Transparency (2017)

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



AI PRINCIPLES in the world – a comprehensive view

Actors:

- Private sector
- Inter-governmental
- Multistakeholder
- Governments
- Civil society

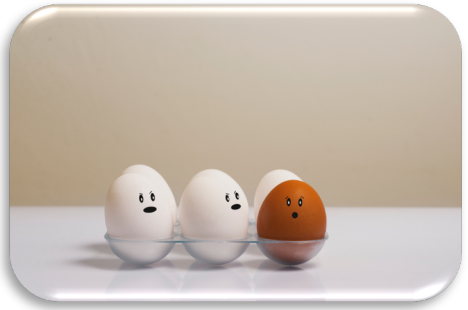
Main themes:

- Human rights
- Human values
- Responsibility
- Human control
- Fairness
- Transparency and explainability
- Safety and Security
- Accountability
- Privacy

Principled AI Project, Berkman Klein's Cyberlaw Clinic, 2019

What does it mean to TRUST a decision made by a machine?

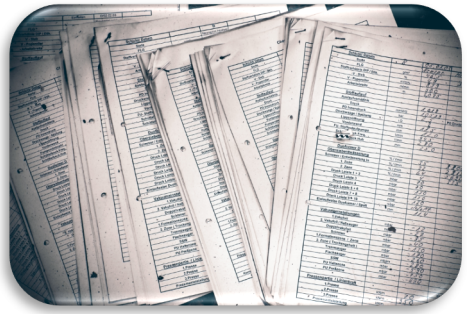
(Other than it is accurate and respect privacy)



Is it **fair**, or is it going to make discriminatory decisions?

Is it possible to understand **why** it made that decision, or is it a black box?

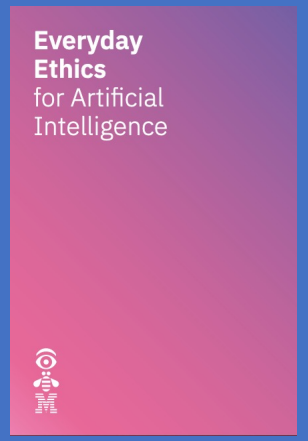
Is it **robust**?

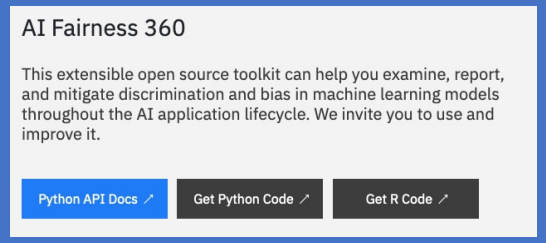


Is it **transparent**?

IBM

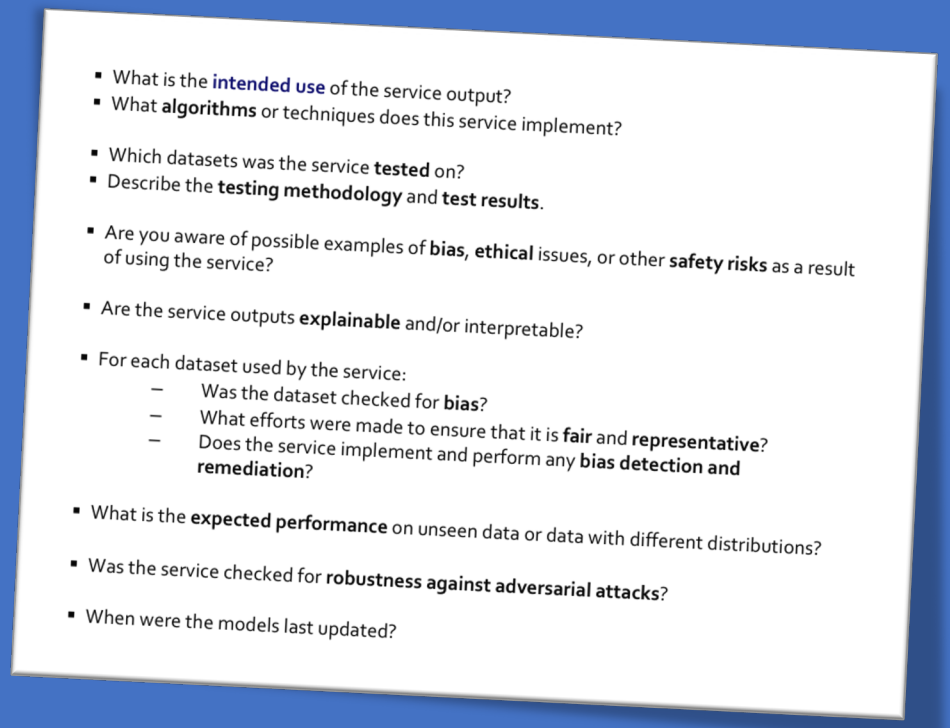
AI fairness at IBM





- Technical solutions to detect and mitigate AI bias
  - Research work
  - Watson OpenScale
  - Open-source libraries: AI fairness 360
- Developers' education and training
  - AI bias education modules for all IBMers
  - Developers' awareness material
  - Revised methodologies for the AI pipeline
  - Adoption strategies
  - Governance frameworks
  - Consultations with all stakeholders
  - Design thinking sessions

AI transparency at IBM

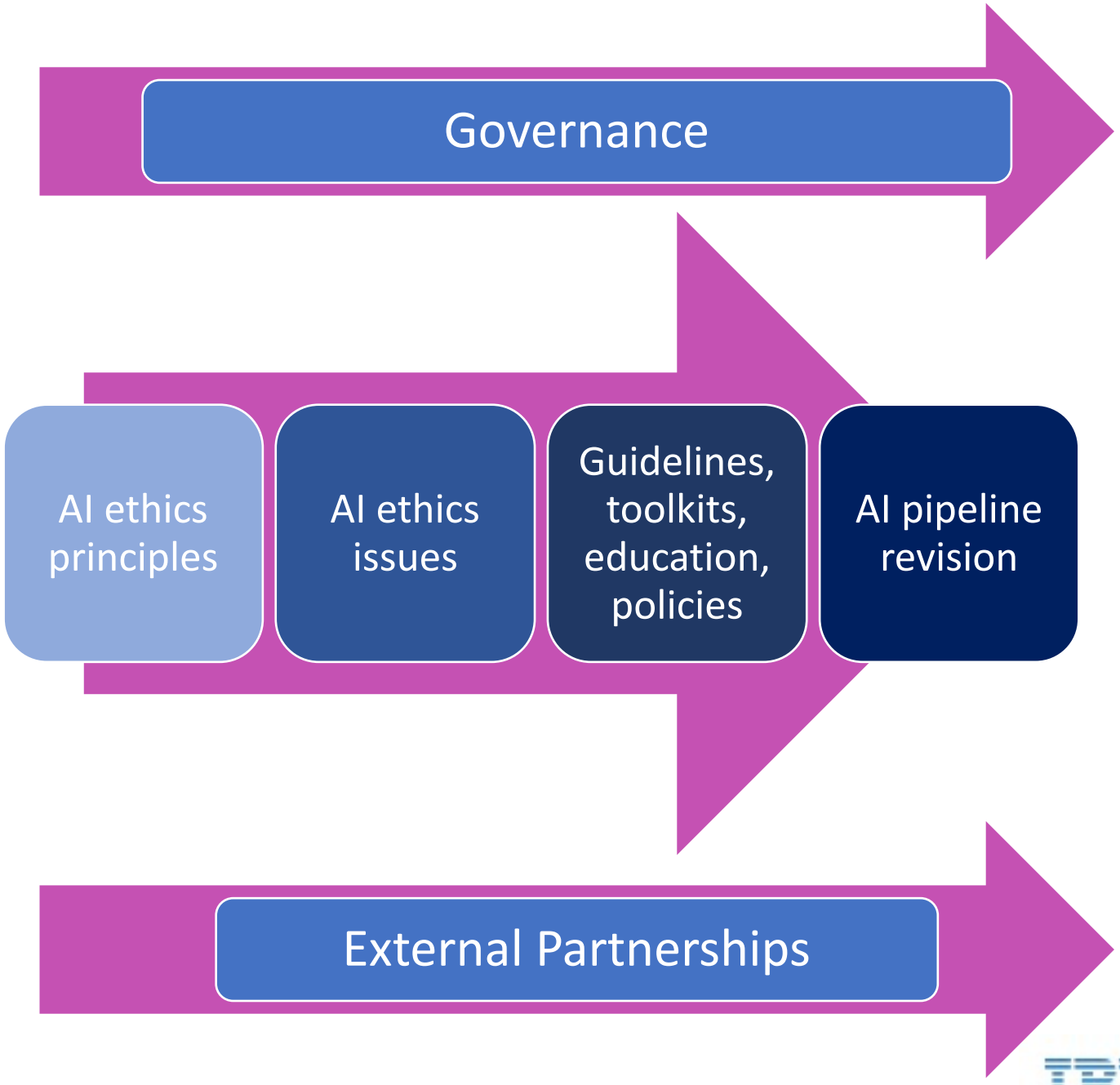


- What is the **intended use** of the service output?
- What **algorithms** or techniques does this service implement?
- Which datasets was the service **tested** on?
- Describe the **testing methodology** and **test results**.
- Are you aware of possible examples of **bias, ethical** issues, or other **safety risks** as a result of using the service?
- Are the service outputs **explainable** and/or interpretable?
- For each dataset used by the service:
  - Was the dataset checked for **bias**?
  - What efforts were made to ensure that it is **fair** and **representative**?
  - Does the service implement and perform any **bias detection and remediation**?
- What is the **expected performance** on unseen data or data with different distributions?
- Was the service checked for **robustness against adversarial attacks**?
- When were the models last updated?

- 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

From principles to practice: a multi-dimensional space



Governance: the IBM AI Ethics board

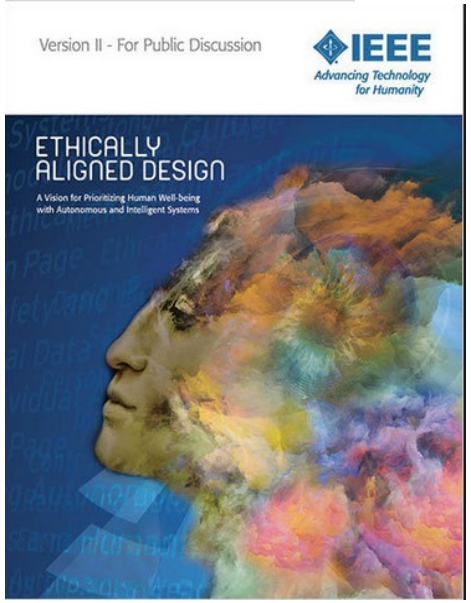
- Mission
  - Awareness and coordination
  - Internal education and retraining
  - Linking research to services and platforms
  - Advice to business units
  - Internal governance framework
  - Define policies and advice regulators
- Risk-based approach for the BUs
  - Vetting based on three dimensions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

Partnerships

Academia
Companies
Governments
Civil society organizations

Multi-disciplinary and multi-stakeholder

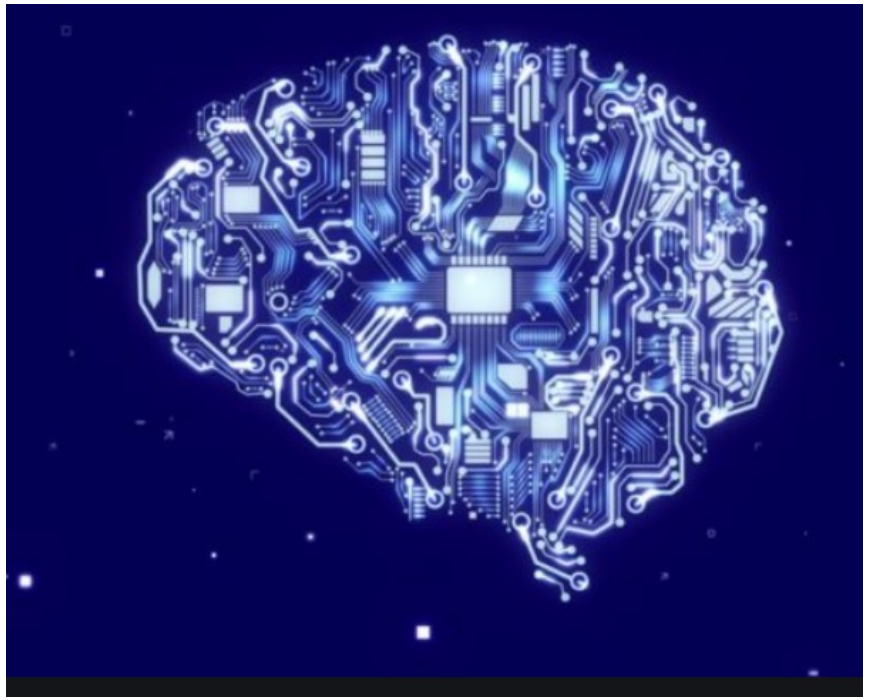
IBM | MIT

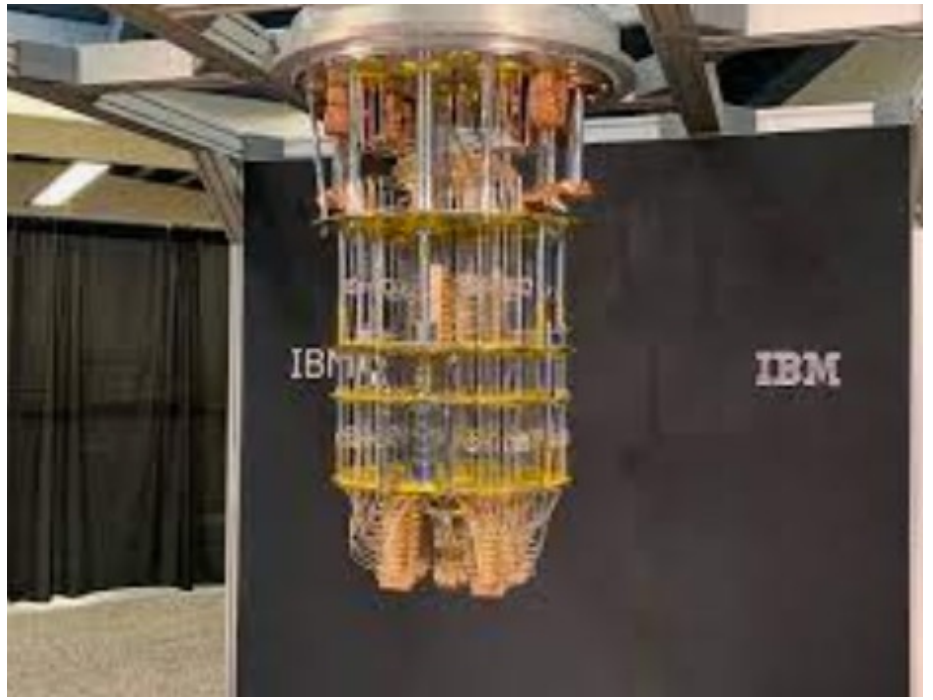


WORLD ECONOMIC FORUM

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?





Useful links

- IBM Approach to AI Ethics:
  - External website: https://www.ibm.com/artificial-intelligence/ethics
  - Trusted AI for business: https://www.ibm.com/watson/ai-ethics/
- Educational material:
  - Everyday Ethics for AI: https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf
- External articles:
  - Harvard Business Review article, 2020: https://hbr.org/2020/11/how-ibm-is-working-toward-a-fairer-ai
- Global studies:
  - IBM IBV study on “Advancing AI ethics beyond compliance”: https://www.ibm.com/thought-leadership/institute-business-value/report/ai-ethics
- Public policies:
  - IBM Policy Lab: https://www.ibm.com/policy/
  - AI precision regulation: https://www.ibm.com/blogs/policy/ai-precision-regulation/
  - Facial recognition: https://www.ibm.com/blogs/policy/facial-recognition/
  - Response to COVID-19: https://www.ibm.com/thought-leadership/covid19/
- Open-source toolkits:
  - AI fairness 360: https://aif360.mybluemix.net/
  - AI explainability 360: https://aix360.mybluemix.net/
  - AI factsheet 360: http://aifs360.mybluemix.net/

IBM

Thank you!

MAI4CAREU | Master programmes in Artificial Intelligence 4 Careers in Europe

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

West Point, April 2021

IBM