Етика на ИИ в IBM: От принципи към практика

Francesca Rossi

Сътрудник на IBM и глобален лидер
по етика на изкуствения интелект

Кратка история на ИИ

ИЗКУСТВЕН ИНТЕЛЕКТ

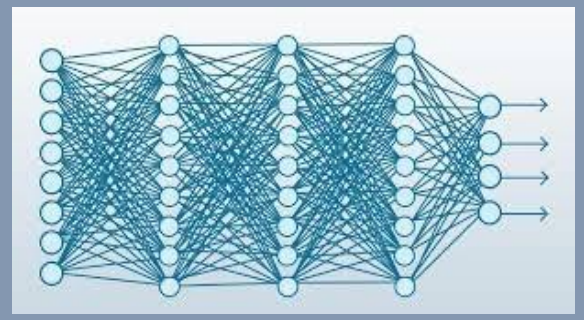
Интелигентни алгоритми, дефинирани и кодирани от хора в машини

МАШИННО ОБУЧЕНИЕ

Способност за учене без изрично програмиране

ДЪЛБОКО УЧЕНЕ

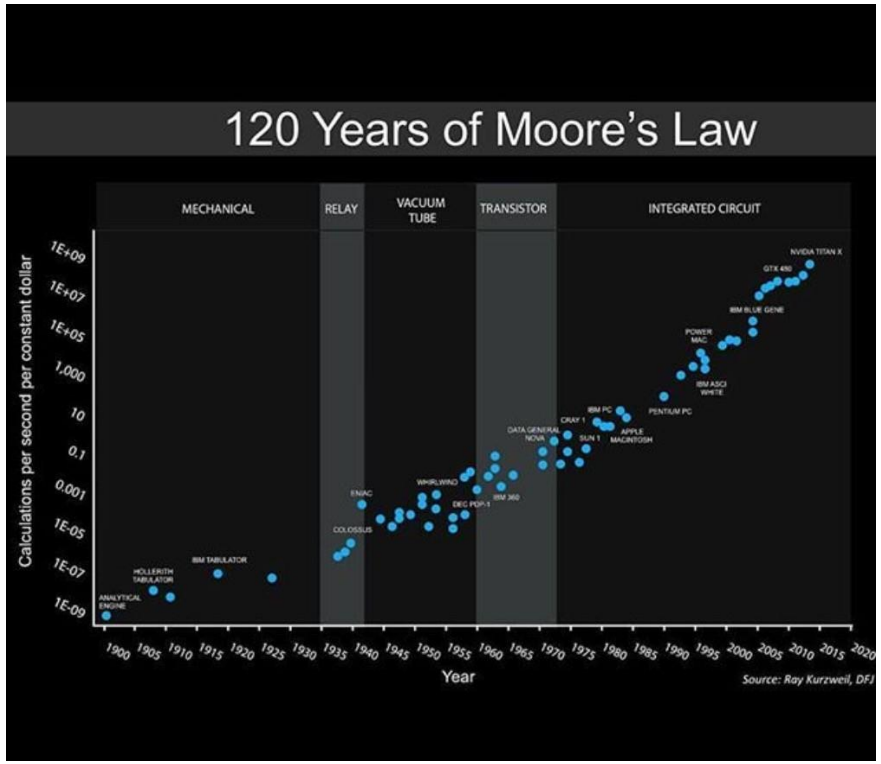
Обучение, базирано на дълбоки невронни мрежи



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

IBM

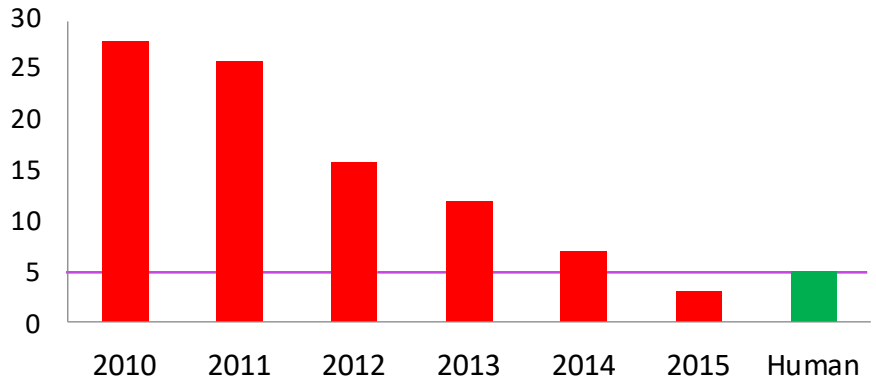
Данни и изчислителна мощност

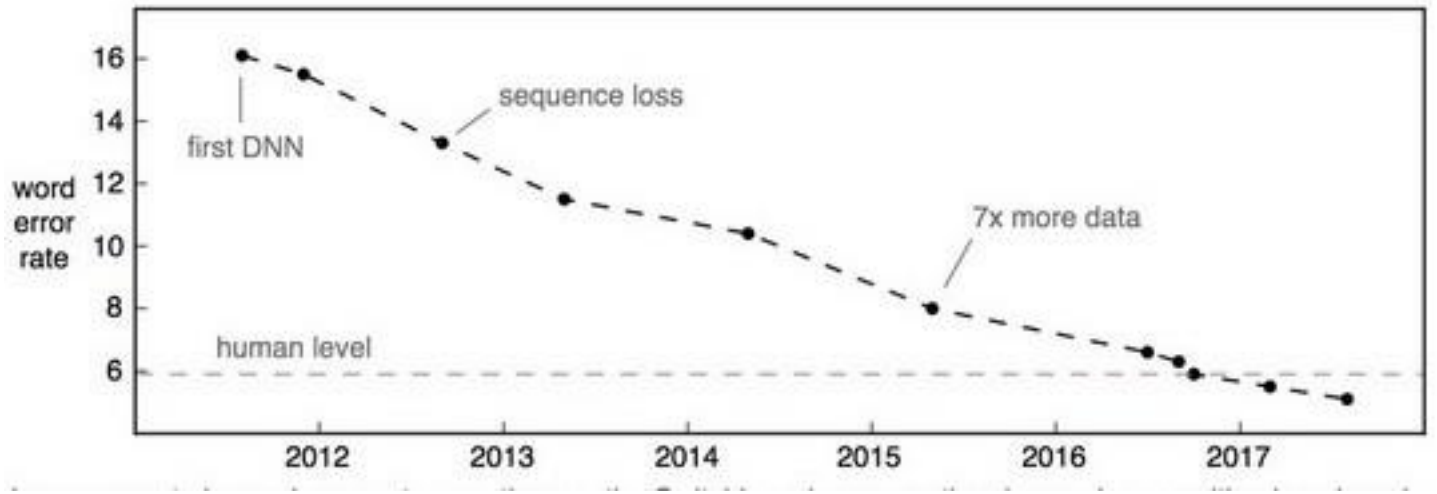


Тълкуване на изображения и естествен език

Жена, която държи бъчва с банани

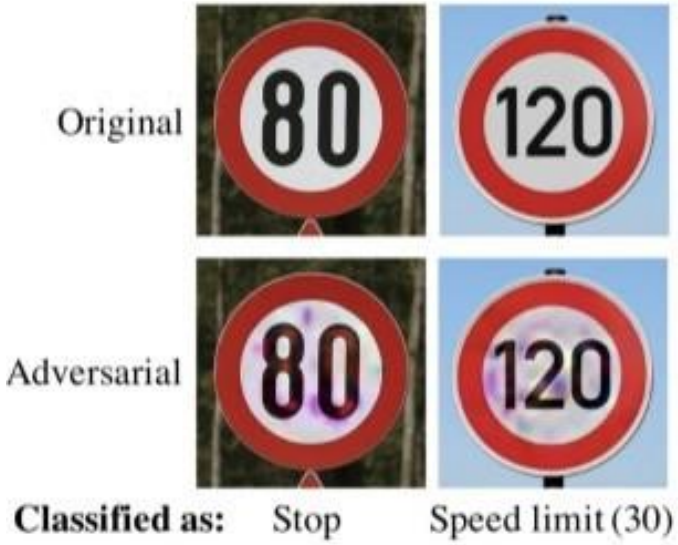
Група млади хора, играещи фризби

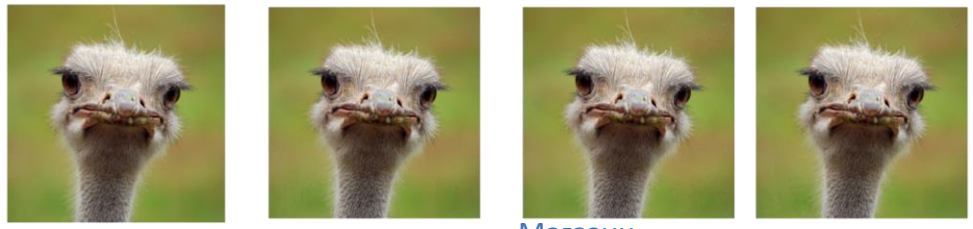




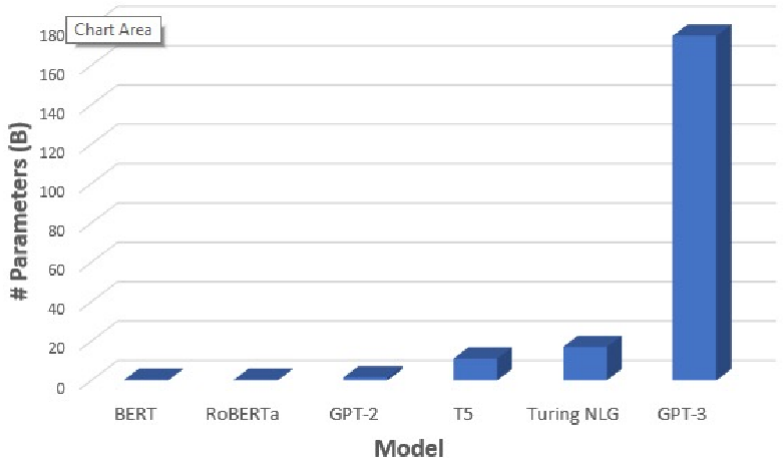
Някои приложения на ИИ

- Дигитални асистенти :
  - Домашни асистенти (Alexa)
  - Асистенти при пътуване (Waze)
- Помощ при шофиране/пътуване :
  - Автопилот (Tesla)
  - Приложения за споделяне на превози (Uber, Lyft)
- Грижа за клиента :
  - Чат ботове за обслужване на клиенти
- Онлайн препоръки :
  - Препоръки на приятели (Facebook)
  - Препоръки за покупка (Amazon)
  - Препоръки за филми (Netflix)
- Медии и новини:
  - Рекламно разположение (Google)
  - Подобряване на новини
- Здравеопазване :
  - Анализ на медицински изображения
  - Препоръка за план за лечение
- Финансови услуги:
  - Оценяване на кредитния риск
  - Одобряване на кредити
  - Откриване на измами
- Пазар на труда :
  - Приоритизиране на автобиография
- Съдебна система :
  - Прогноза за рецидивизъм (Compas)





Щраус безопасно Магазин за обувки Вакуум



Ограничения на ИИ

- Тесен ИИ
  - Решава добре специфични проблеми
- Липса на устойчивост и адаптивност
- Нуждае се от много ресурси
  - Данни и изчислителна мощност

IBM

Етични въпроси - примери

Зависим от пола процес на одобрение на кредитна карта на Apple

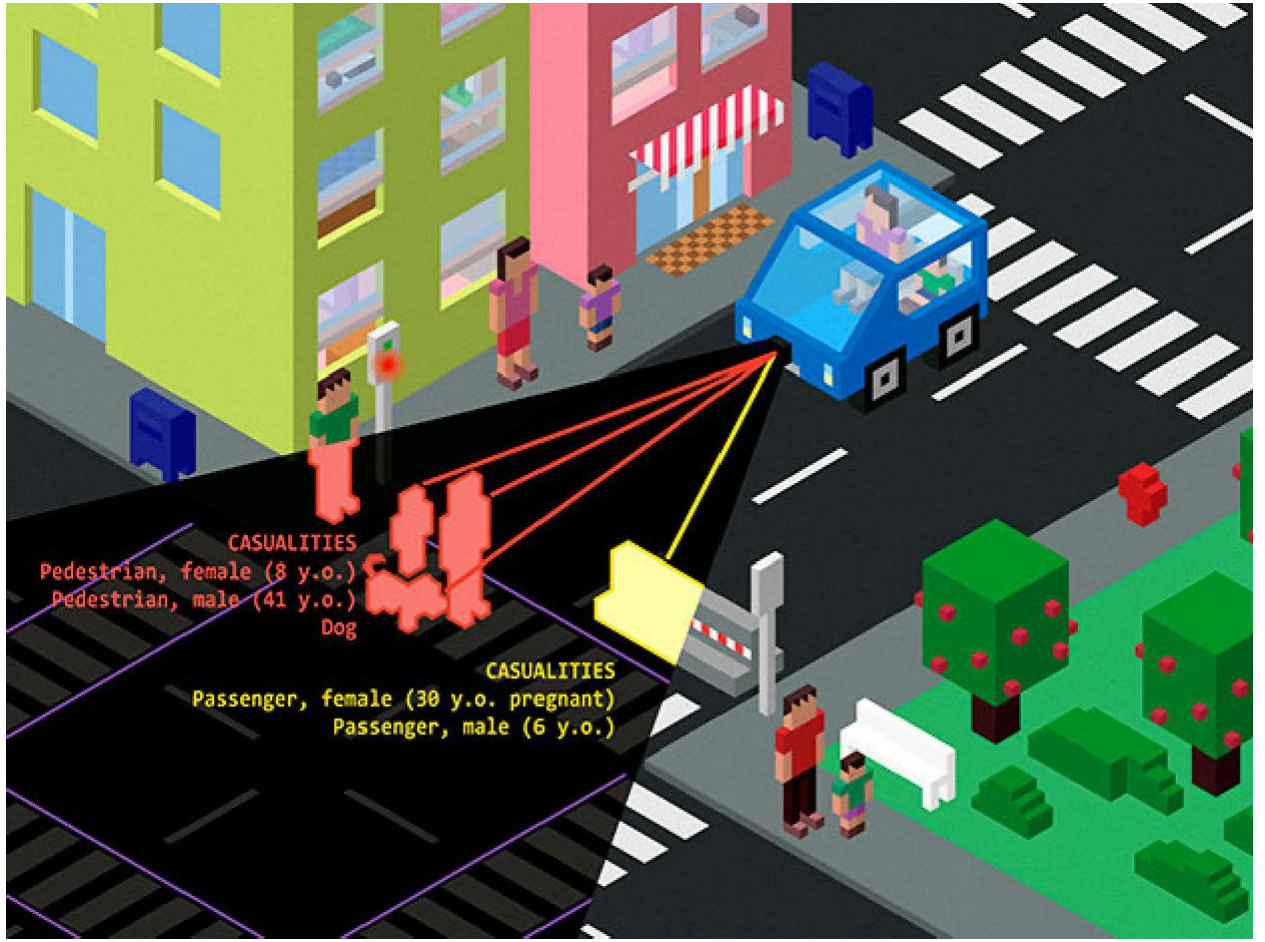
Дискриминация при динамичното ценообразуване на споделеното пътуване

софтуер за набиране на персонал взависимост от пола

Чатбот, който демонстрира расистки изказвания

Неетично използване на лични данни

Можем ли да се доверим на решенията на ИИ?



Етика на ИИ

Мултидисциплинарна област на обучение

Как да оптимизираме благоприятното въздействие на ИИ, като същевременно намалим рисковете и неблагоприятните резултати

Как да проектираме и изградим системи с ИИ, които са наясно с ценностите и принципите, които трябва да се следват в сценариите за внедряване

Да идентифицираме, проучим и предложим технически и нетехнически решения за етични проблеми, произтичащи от широко разпространеното използване на ИИ в живота и обществото

Основни въпроси, свързани с етиката на ИИ

ИИ се нуждае от данни

- Поверителност на данните и управление

ИИ често е черна кутия

- Обяснимост и прозрачност

ИИ може да взема или препоръчва решения

- Справедливост и ценностно изравняване

ИИ е базиран на статистика и има малък процент грешка

- Кой носи отговорност, ако станат грешки?

ИИ може да профилира хората и да манипулира техните предпочитания

- Човешка и морална свобода на действие

ИИ е много широко разпространен и динамичен

- По-големи отрицателни въздействия за злоупотреба с технологии
- Бърза трансформация на работни места и общество

Добро или лошо използване на технологията

- Автономни оръжия и масово наблюдение
- Цели на ООН за устойчиво развитие

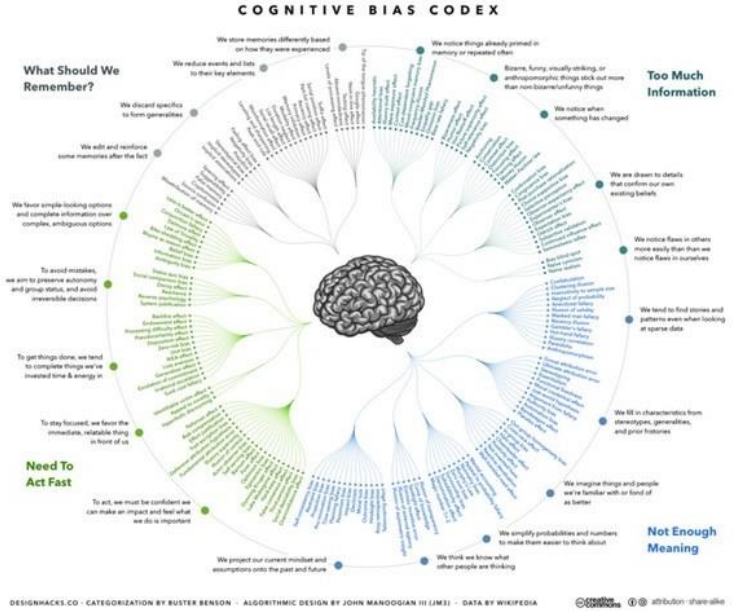
IBM

ИИ не е неутрална технология

- Трябва да се избягва злоупотреба
- Но ИИ трябва да бъде проектиран и разработен с правилните свойства
  - Честно, обяснимо, стабилно, …

Справедливост на ИИ

- Пристрастие: предубеждение за или против нещо
- Като следствие от пристрастие, човек може да се държи несправедливо към определени групи в сравнение с други
- Защо AI трябва да е предубеден?
  - Обучен с данни, предоставени от хора, а хората са предубедени



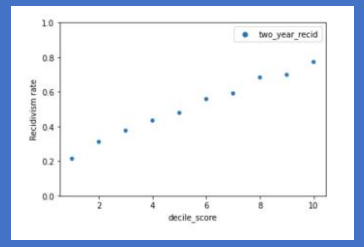
Пристрастие на ИИ ImageNet

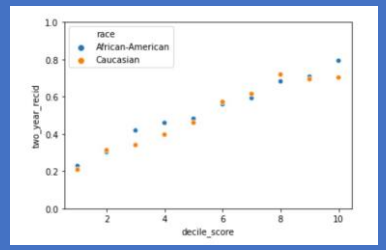
- 14M изображения, използвани за обучение на AI системи за интерпретация на изображения
- Пристрастия в разпространението на данни и в етикетите на данните(Mturk people)

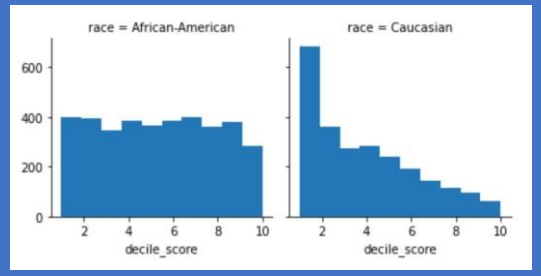
Молба за ипотека: пристрастие не само от данни

- Данни за обучение
  - Пр. : корелация пол-приемане
- Решение проектиране:
  - Пр.: приоритетни мотиви за кандидатстване за заем
  - Купуване на къща
  - Плащане на училищни такси
  - Плащане на адвокатски такси
    - Молбите за заем с тези мотиви са приоритетни
    - Ако един от тях е пропуснат, съответната общност ще бъде санкционирана

Пристрастие на ИИ : което е правилното определение за справедливост?







- Общата точност е една и съща, независимо от расата (**равенство на обща точност**)
- Вероятността от рецидив сред обвиняемите, обозначени като среден или висок риск, е сходна, независимо от расата (**предсказуем паритет**)
- Но … фалшивите положителни и фалшиво отрицателните нива са много различни

	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

Много точки при вземане на решения

- ***Индивидуална срещу групова справедливост:***
  - подобни индивиди трябва да получават подобно отношение или резултати, срещу
  - групите, дефинирани чрез защитени атрибути, трябва да получават подобно отношение или резултати
- ***Зависещи от контекста дефиниции за справедливост***
- **Приемлив праг на пристрастие**
- ***Кога да се открие пристрастие:***
  - данни за обучение или обучен модел

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

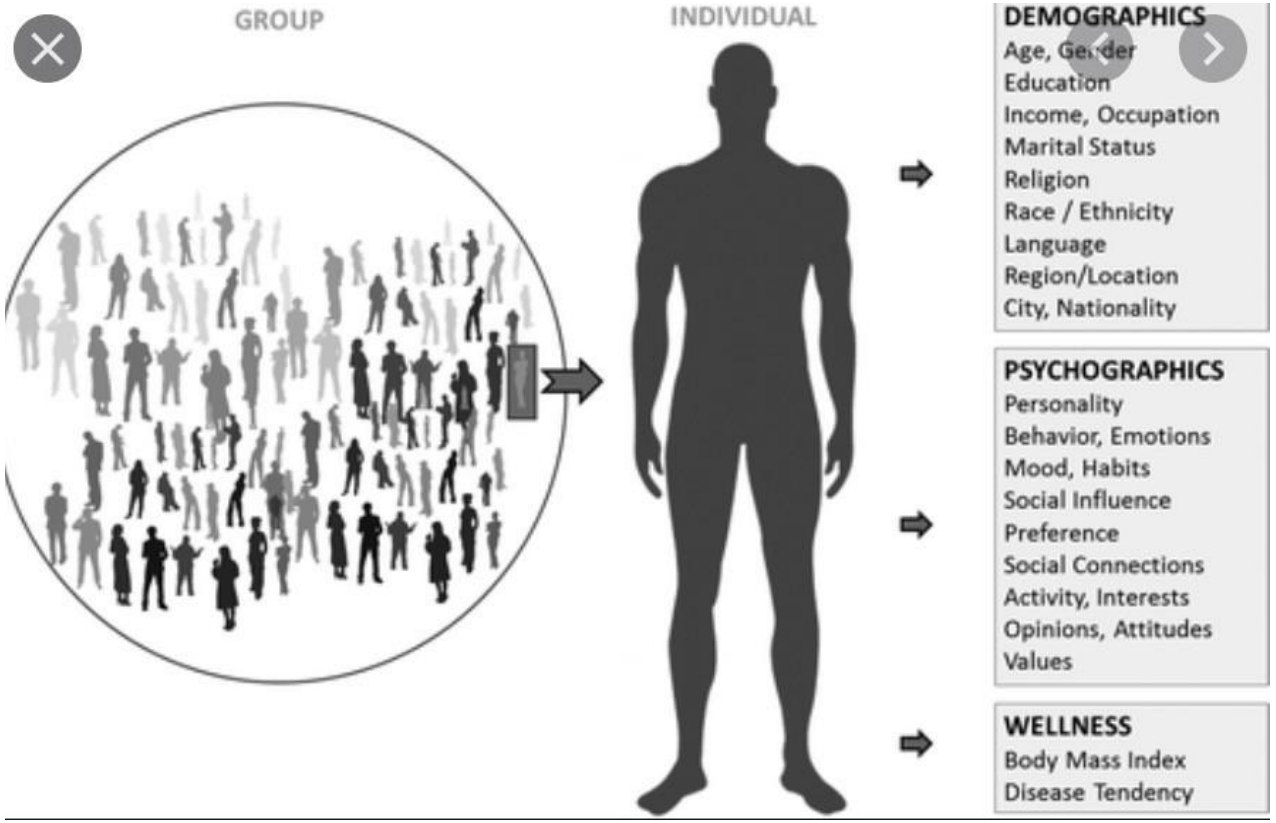
Name	Closest relative	Note	Reference
Statistical parity	Independence	Equivalent	Dwork et al. (2011)
Group fairness	Independence	Equivalent	
Demographic parity	Independence	Equivalent	
Conditional statistical parity	Independence	Relaxation	Corbett-Davies et al. (2017)
Darlington criterion (4)	Independence	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)

Обяснимост на ИИ: системите с ИИ не могат да бъдат черни кутии

Обработка на данни: Общият регламент за защита на данните (GDPR)

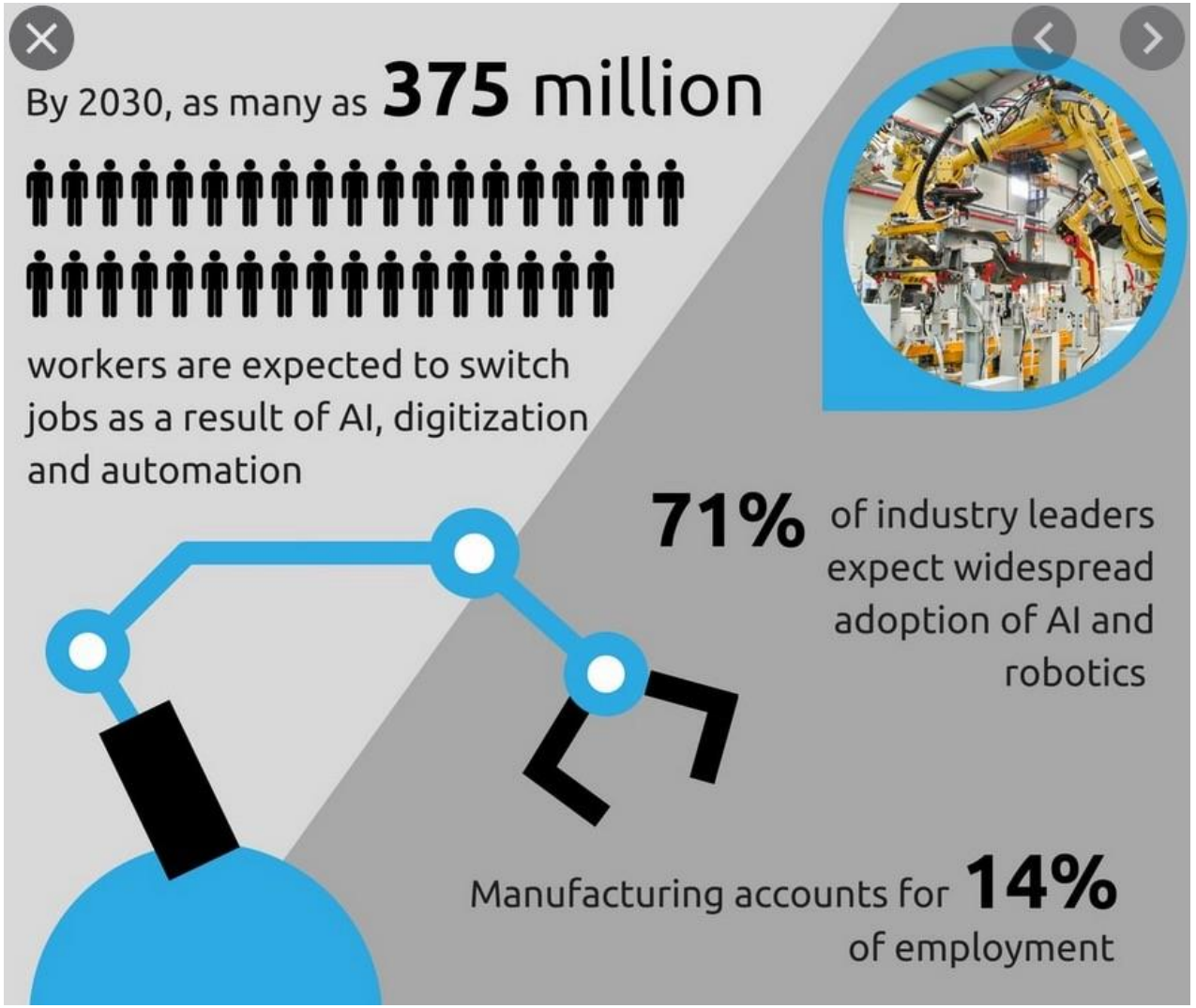
Профилиране и манипулация

- От действия до профили
  - Харесвам, текст, изображения, следвам, …
- AI може да открие нашите предпочитания и да ги използва, за да рекламира продукти, които вероятно харесваме
  - По-лесно, ако предпочитанията ни са биполярни



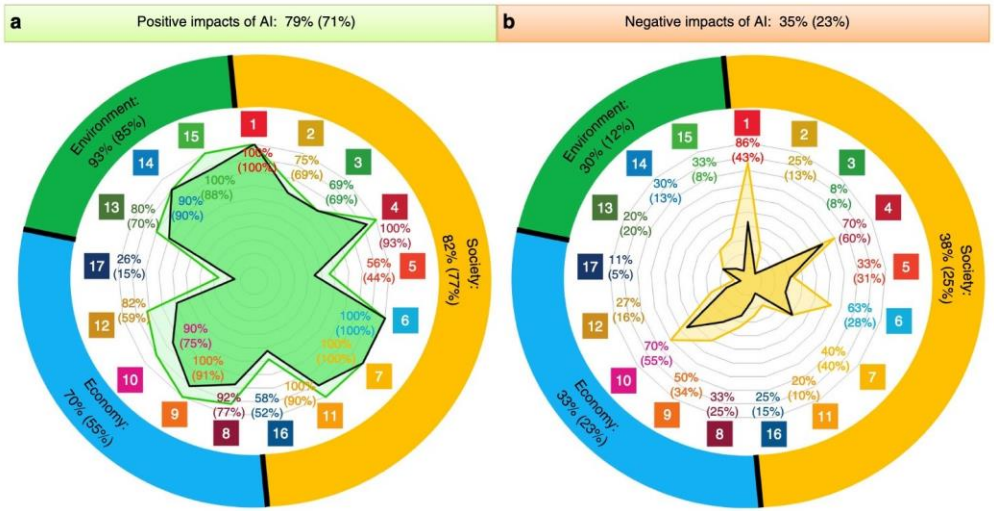
Въздействие върху работната сила

- Много работни места ще изчезнат и много други ще бъдат създадени
- Всички работни места ще се променят



Визия за бъдещето (2030)

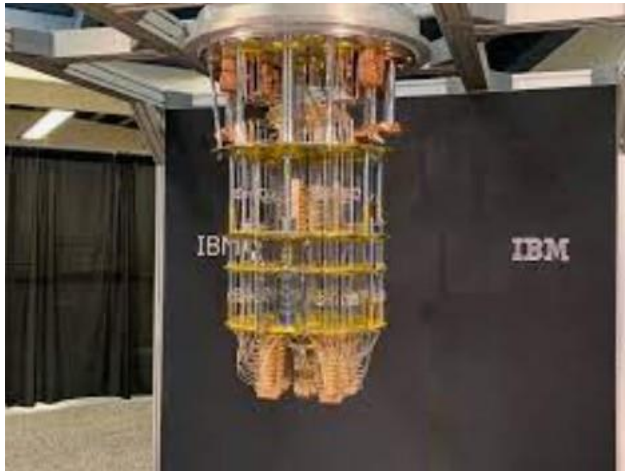
- 17 цели на устойчиво развитие
- Много труден път
  - Пандемията влоши ситуацията
- ИИ може да помогне при постигане на целите
- COVID: ваксини!



IBM, технологии и ИИ

- 110 години
- Хардуер и софтуер
- ИИ решения за други компании
  - Банки и финансови институции
  - Правителства
  - Летища
  - Болници
  - …

Summit, IBM



Quantum computer, IBM

Chess: IBM Deep Blue, 1997

Jeopardy: IBM Watson, 2011

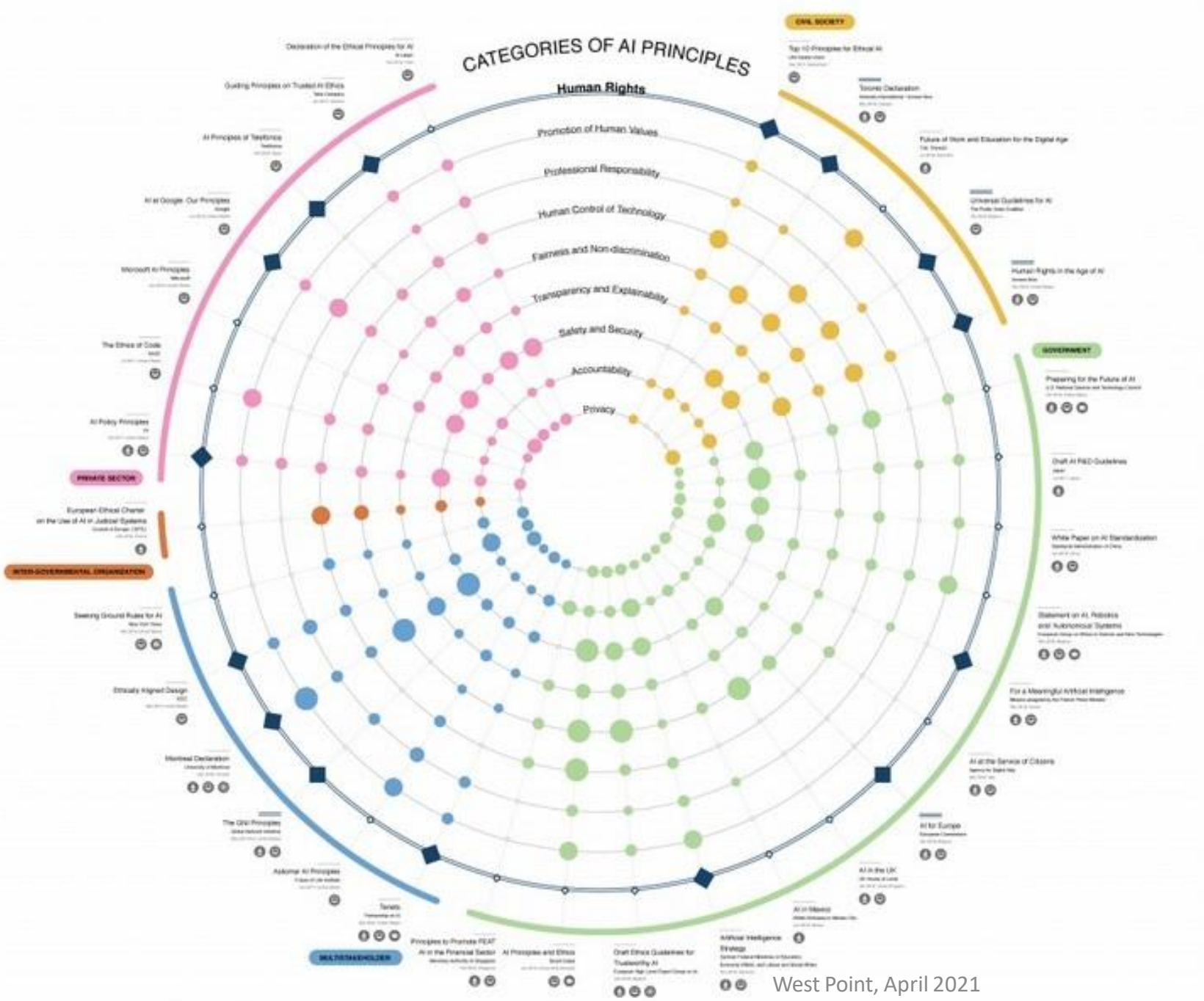
Project Debater, 2020

Принципи на IBM за доверие и прозрачност (2017)

Целта на ИИ е да увеличи човешкия интелект

Данните и прозренията принадлежат на техния създател

Новите технологии, включително системите с ИИ, трябва да бъдат прозрачни и обясними



ПРИНЦИПИТЕ НА ИИ в света – цялостен поглед

Актьори:

- Частен сектор
- Междудържавни
- Множество заинтересовани страни
- Правителства
- Гражданско общество

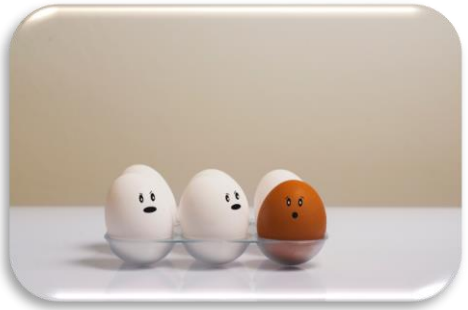
Основни теми:

- Човешки права
- Човешки ценности
- Отговорност
- Човешки контрол
- Справедливост
- Прозрачност и обяснимост
- Безопасност и сигурност
- Отчетност
- Поверителност

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

Какво означава да се ДОВЕРИШ на решение, взето от машина?

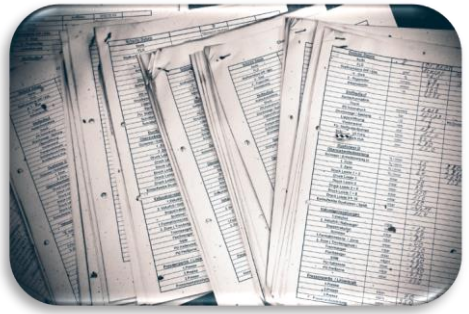
(Освен това да е точно и да зачита поверителността)



Справедливо ли е, или ще взема дискриминационни решения?

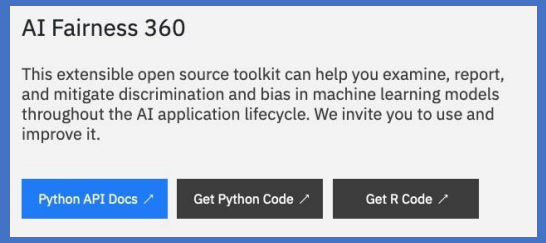
Възможно ли е да се разбере **защо** е взето това решение или е черна кутия?

Устойчиво ли е?



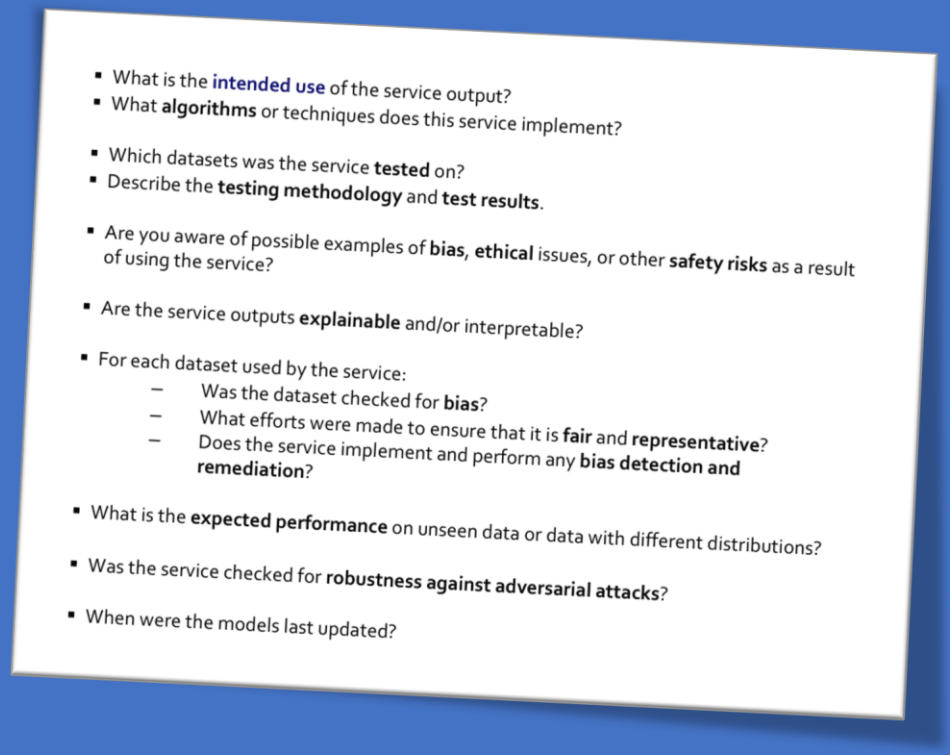
Прозрачно ли е?

Справедливост на ИИ в IBM



- Технически решения за откриване и смекчаване на пристрастията на ИИ
  - Изследователска работа
  - Watson OpenScale
  - Библиотеки с отворен код: AI fairness 360
- Образование и обучение на разработчици
  - Образователни модули за пристрастия на ИИ за всички в IBM
  - Информационен материал за разработчиците
  - Ревизирани методологии за ИИ
  - Стратегии за осиновяване
  - Рамки за управление
  - Консултации с всички заинтересовани страни
  - Сесии за дизайнерско мислене

Прозрачност на ИИ в 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?

- Информационен лист за ИИ
  - Прозрачност чрез документация
  - Проектиране на избор за развитие
  - Не просто контролен списък
  - Самооценка и не само
- Полезен за
  - Разработчици
  - Клиенти
  - Потребителски регулатори/одитори
- В съответствие с Експертната група на високо ниво на ЕК относно списъка за самооценка на ИИ (ALTAI)
- AI factsheet 360

IBM

От принципи към практика: многомерно пространство

Управление: IBM борд по етика на ИИ

- Мисия
  - Съзнателност и координация
  - Вътрешно обучение и преквалификация
  - Свързване на изследвания с услуги и платформи
  - Съвети към бизнес единици
  - Рамка за вътрешно управление
  - Дефиниране на политики и съвети за регулаторите
- Основан на риска подход за BU
  - Проверка въз основа на три измерения (технология, употреба, клиент)

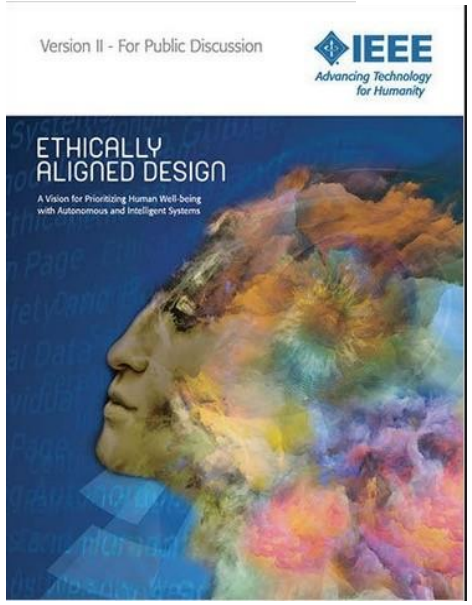
Партньорства

Академични среди
Компании
Правителства
Организации на гражданското общество
Мултидисциплинарни и много заинтересовани страни

IBM | MIT

AAAI / ACM conference on ARTIFICIAL INTELLIGENCE, ETHICS, AND SOCIETY

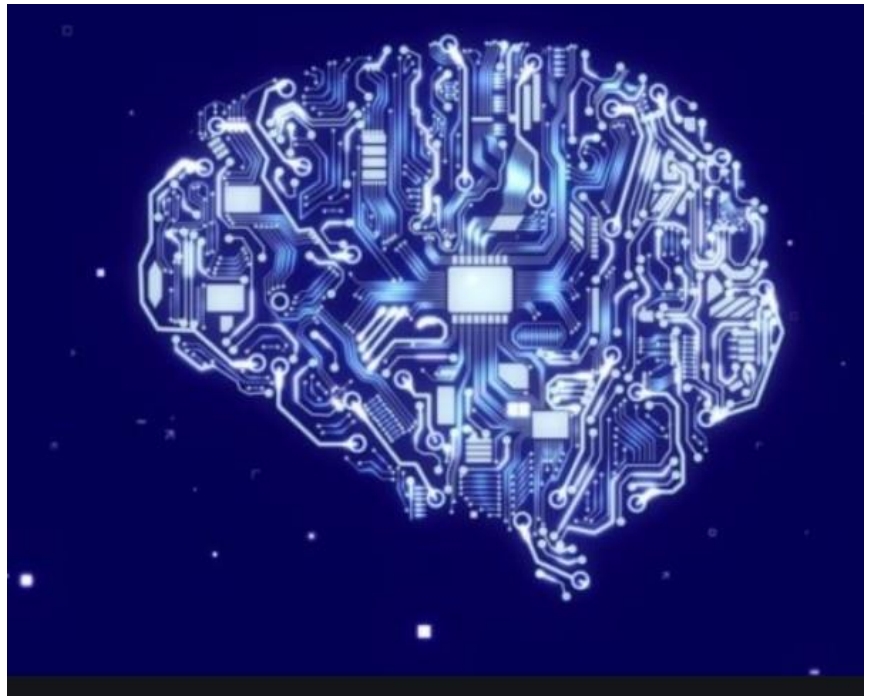
AI for Good Global Summit
An ITU experience

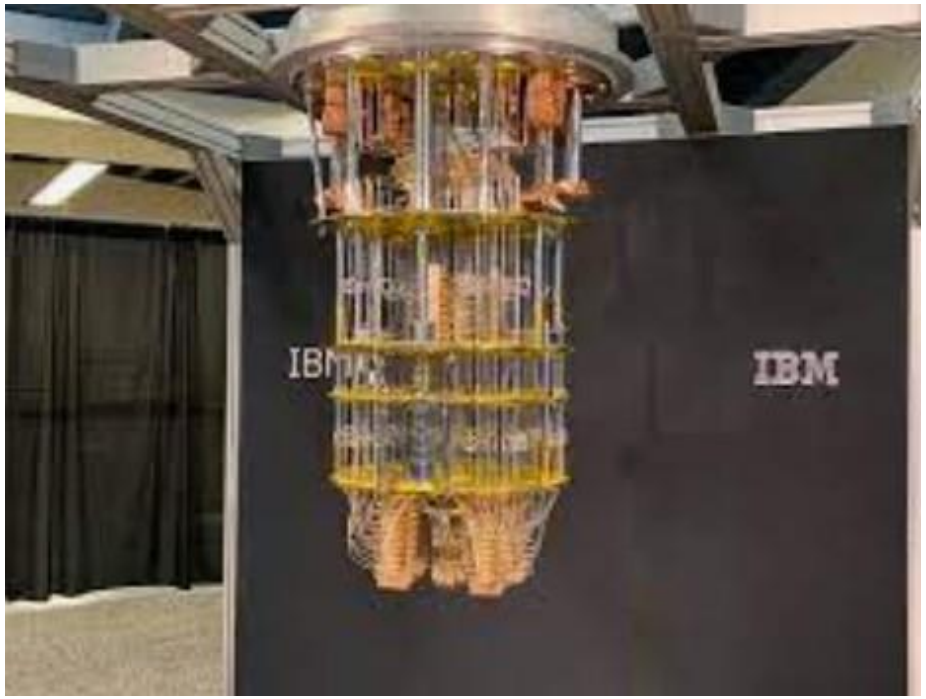


WORLD ECONOMIC FORUM

Не просто ИИ

- Невротехнологии
  - Огромен потенциал за здравеопазване
  - Четене/запис на невроданни
  - Допълнителни проблеми около поверителността, агенцията и самоличността
- Квантови изчисления
  - Как да използваме отговорно такава огромна изчислителна мощност?





Полезни връзки

- Подход на IBM към етика в ИИ:
  - External website: https://www.ibm.com/artificial-intelligence/ethics
  - Trusted AI for business: https://www.ibm.com/watson/ai-ethics/
- Образователен материал:
  - Everyday Ethics for AI: https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf
- Външни статии:
  - Harvard Business Review article, 2020: https://hbr.org/2020/11/how-ibm-is-working-toward-a-fairer-ai
- Глобални изследвания :
  - IBM IBV study on “Advancing AI ethics beyond compliance”: https://www.ibm.com/thought-leadership/institute-business-value/report/ai-ethics
- Публични политики:
  - 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/
- Инструментариуми с отворен код:
  - AI fairness 360: https://aif360.mybluemix.net/
  - AI explainability 360: https://aix360.mybluemix.net/
  - AI factsheet 360: http://aifs360.mybluemix.net/

Благодаря!