Lecture 10.2 - The Ethics of AI

Gleb Sizov
Norwegian University of Science and Technology

> Given that AI is a powerful technology, we have a moral obligation to use it well, to promote the positive aspects and avoid or mitigate the negative ones

Positive side of AI

1. AI in crop management and food production help feed the world
2. AI in drug helps to cure diseases
3. Drive assistance helps make driving safer
4. Optimization of business processes increases wealth
5. Automation for tedious and dangerous tasks
6. AI-based assistance for people with disabilities
7. Smart grid and buildings save energy and infrastructure
8. Machine translation helps people to communicate

   ...

Principles of Robotics

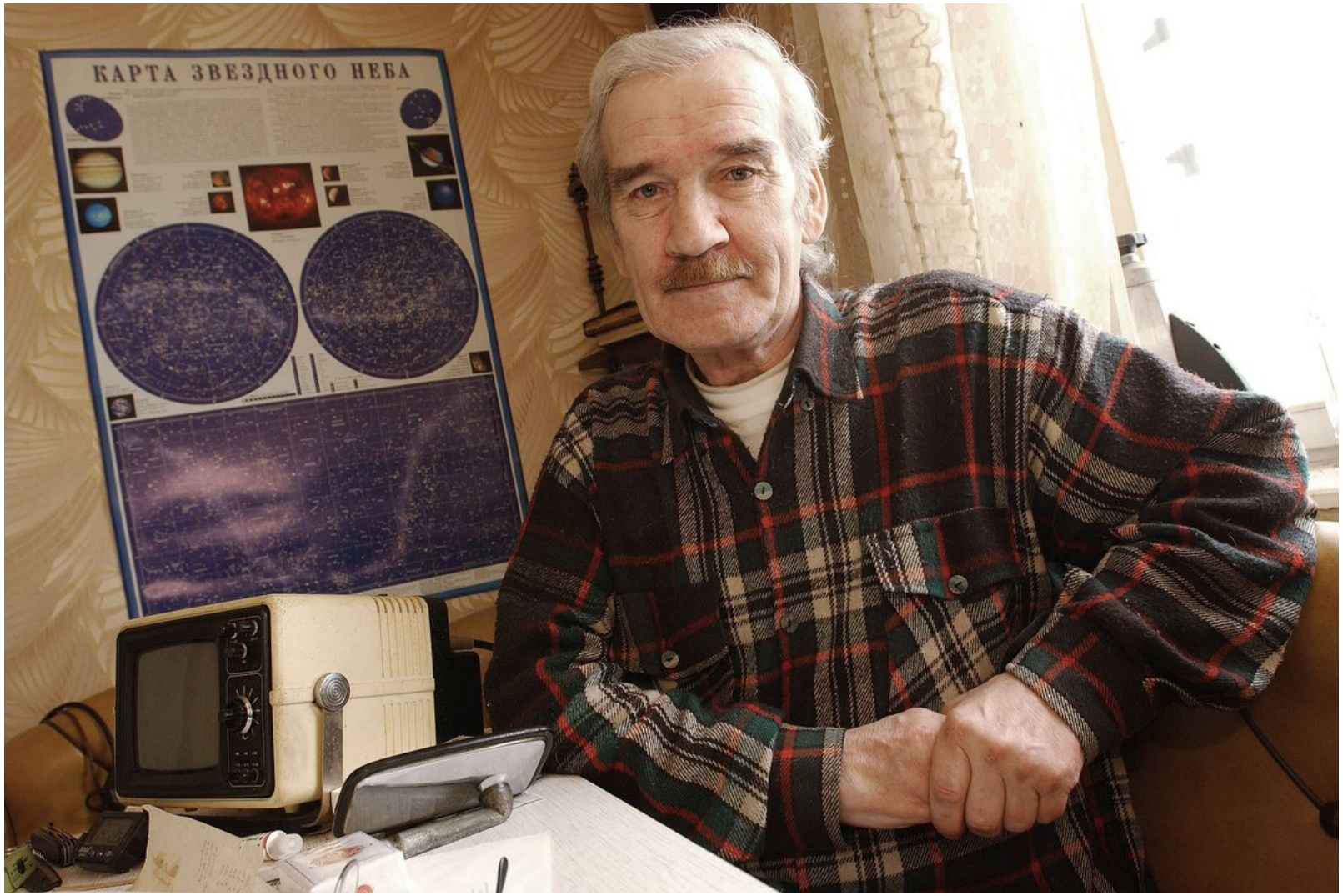
Ensure safety	Establish accountability
Ensure fairness	Uphold human rights and values
Respect privacy	Reflect diversity/inclusion
Promote collaboration	Avoid concentration of power
Provide transparency	Acknowledge legal/policy implications
Limit harmful uses of AI	Contemplate implications for employment

UK’s Engineering and Physical Sciences Research Council, 2010

Lethal autonomous weapons - Issues

1. Discriminating between combatants and non-combatants
2. Judgment of attack necessity
3. Assesment of potential collateral damage
4. Reliability
5. Cybersecurity
6. Scalable weapons of mass destruction
7. Attacker advantage

On September 26, 1983, Stanislav Petrov saved the world



Lethal autonomous weapons - Regulations, treaties

- 2014 - Convention on Certain Conventional Weapons (CCW)
  Ban of lethal autonomous weapons
  Supported by 30 nations, opposed by Israel, Russia, US
- 2015 - **Compaign to Stop Killer Robots**
  140 NGOs in over 60 countries
  4000 AI researchers and 22000 others

> AI is a **dual use** technology

Surveillance

1976 - Joseph Weizenbaum warned that automated speech recognition could lead to widespread wiretapping, and hence to a loss of civil liberties

2018 - 350 million surveillance cameras in China and 70 million in US.

2019 - China's social credit system puts emphasis on use of big data and AI

Privacy

1. Data collectors have a moral and legal responsibility to be good stewards of the data they hold.
2. Balance between privancy rights and sociaety gains from sharing data, e.g. stop terrorists and cure diseases.

Regulations:

US: Privacy of medical and student records:

- Health Insurance Portability and Accountability Act (HIPAA)
- Family Educational Rights and Privacy Act (FERPA)

EU: General Data Protection Regulation (GDPR)

Privacy protection methods

1. De-identification, e.g. remove field, can be re-identified
2. Generalizing fields, e.g. 20-30 years old
3. K-anonimity - indistinguishable form at least k - 1 records
4. Aggregate quering - rules for minimum number of items
5. Differential privacy - add noise to query results
6. Federated learning - share model parameters, not data

Security

Market for ML cybersecurity $100 billion by 2021

- *Attackers* use AI to automate **cyberscurity** attacks.
- *Defenders* use AI to detect anomlaous network traffic and fraud transactions

Designing secure AI systems:

1. High compelxity, autnonomy, dependance on data
2. Not a well established field

Fairness and bias

> Designers of machine learning systems have a moral responsibility to ensure that their systems are in fact fair.

Avoid discrimination by algorithms, bias in data.

Domains:

- Loan approaval
- Education
- Employment
- Housing

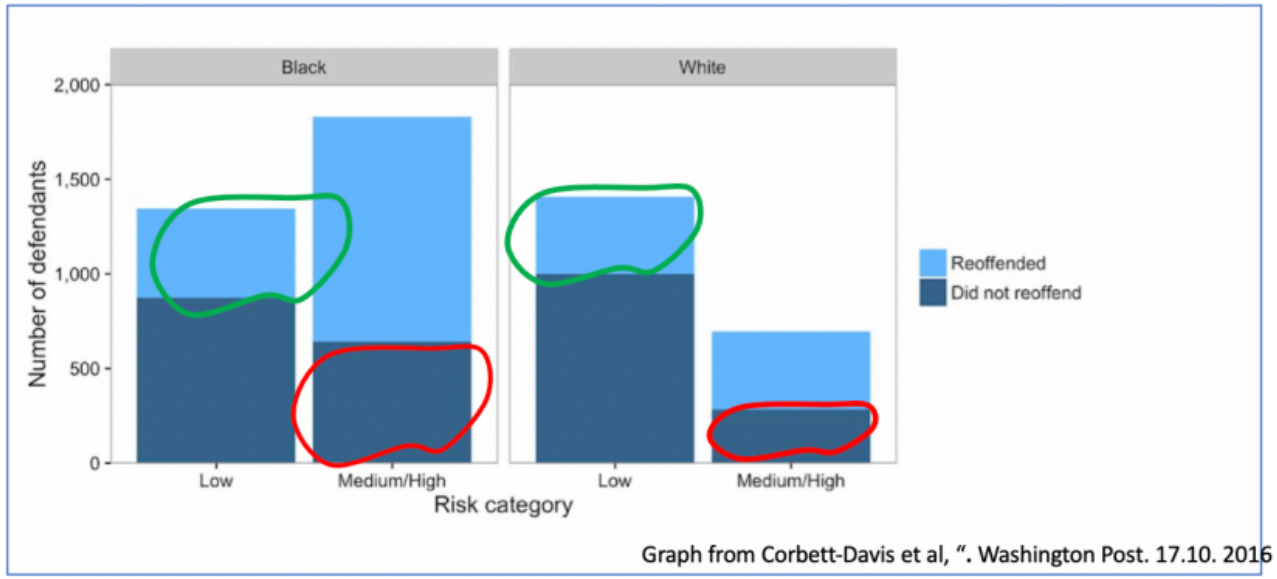
Fairness

1. Individual fairness - similar individuals are treated similarly
2. Group fairness - two classes are treated similarly
3. Fainess through unawareness - e.g. remove gender and race fields
4. Equal outcome - each demographic class get the same result
5. Equal opportunity - individuals are treated according to their true ability, regardless of the class.
6. Equal impact - individuals with similar ability should have the same expected utility, regardless of the class

COMPAS example

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

ProPublica argues: COMPAS is not fair



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

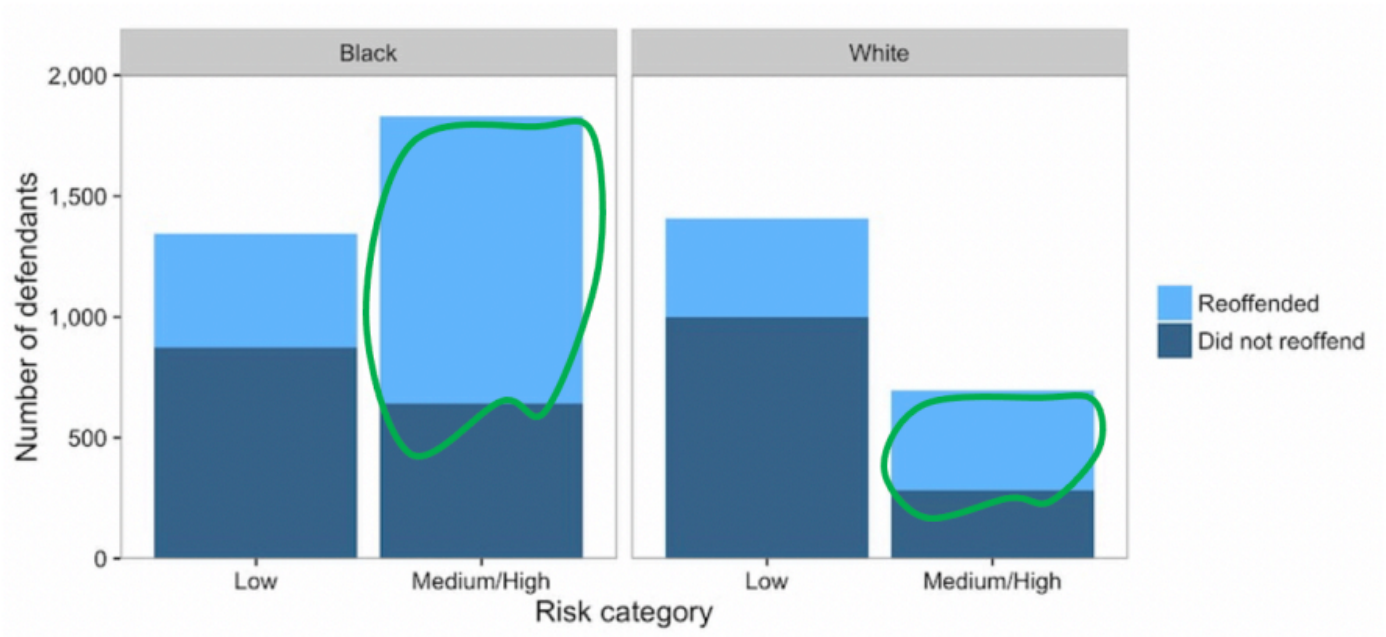
	predicted high risk	predicted low risk
reoffended	True. Positive.	False Negative
not reoffended	False. Positive.	True Negative

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

Northpointe argues: COMPAS is fair

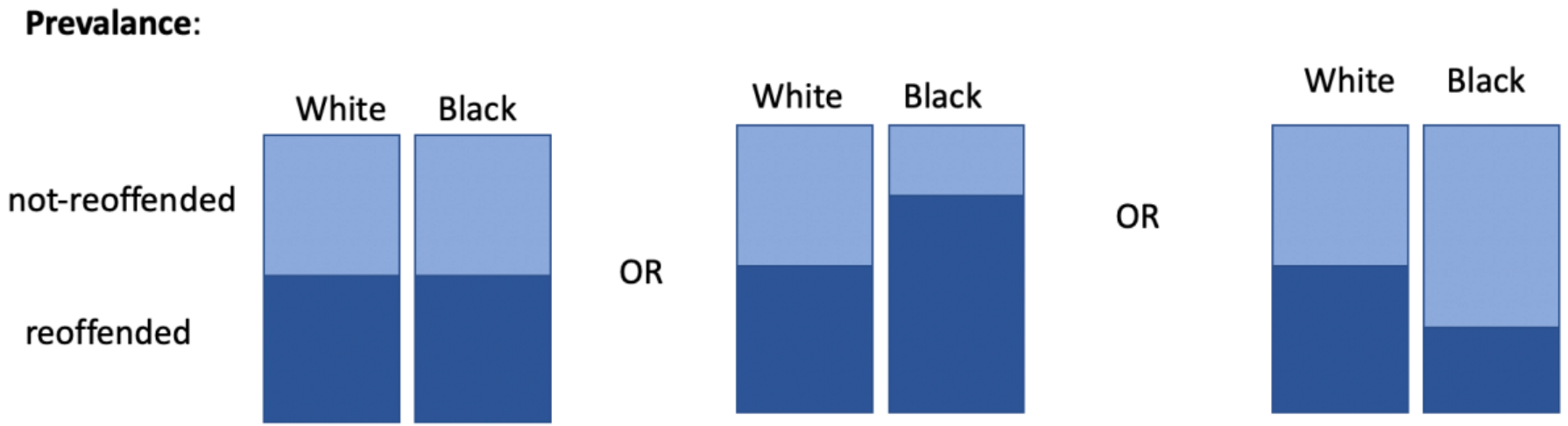
- COMPAS is fair because: among defendants that COMPAS predicted as “highly probable to reoffend”:
  - 60 % of the white defendants actually did reoffended
  - 61 % of the black defendents reoffended

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



Are both fairness goals possible simultaneously?

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



Fairness and bias issues

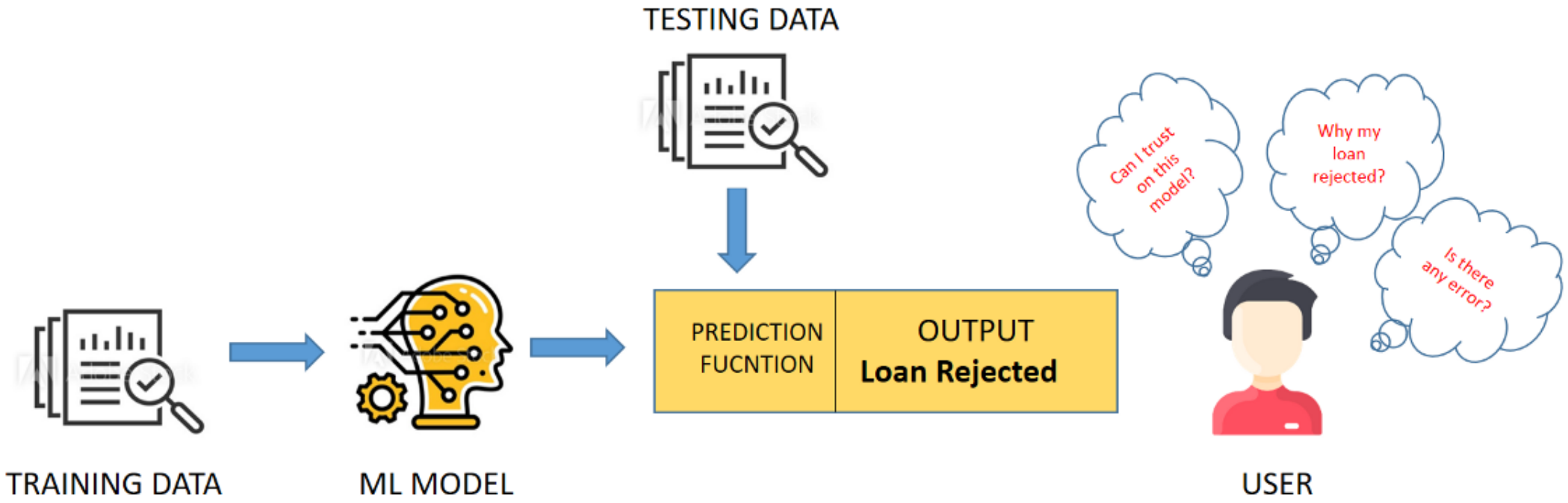
1. Biased data generated by biased sociatal process.
2. AI is used to *justify* bias.
3. Finding the right (fair) objective, e.g. exisitng skills vs learning on the job.
4. Decide which classes to protect, e.g. race, color, religion, sec, ddisability, familyt status etc.
5. Samples size disparity
6. Bias in software development process

Defences against bias

1. Understand the limits of your data, provide model data sheets
2. Diversity among engineers
3. De-bias your data, e.g. over-sampling.
4. ML algorithms that more resistant to bias.
5. Monitor metrics for different groups and fix bias issues.

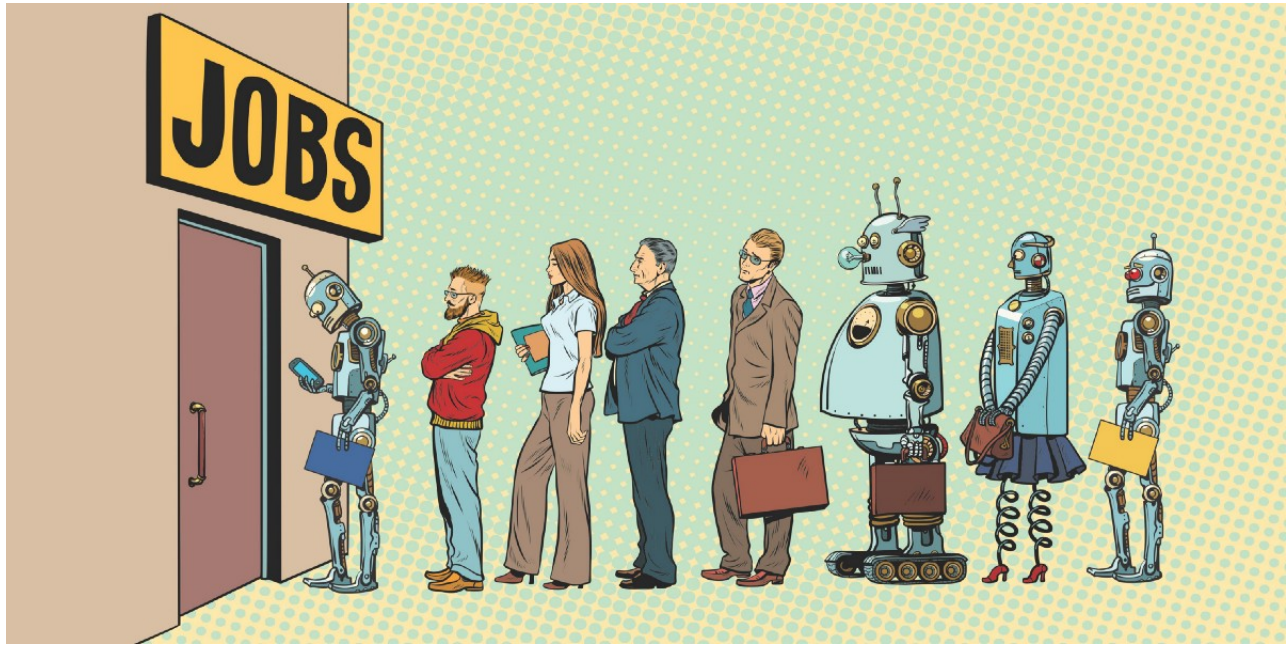
Trust and transperancy

1. Verification and validation methodology
2. Certification of AI systems
3. Transparency - explainable AI



The future of work

1. Technological unemployment
2. Buisness process automation
3. Income inequality

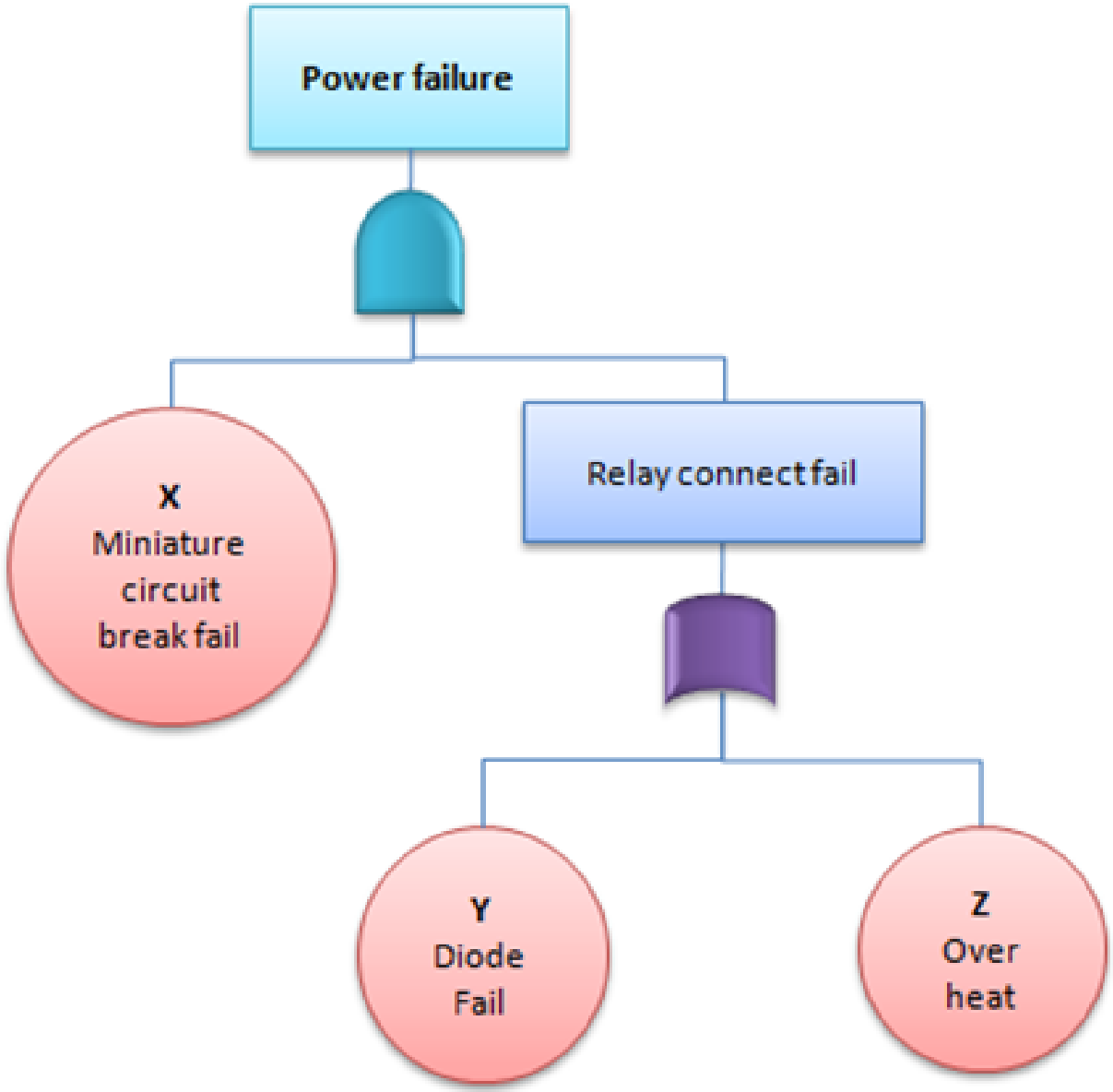


AI Safety

Safety-critical applications:
Driving, robotics, construction, medicine

Safety engineering

1. Failure modes (FMEA) - analyze what can go wrong
2. Falut tree analysis
3. Software engineering practices
4. Utility function analysis - side effects, value alignment



Ultraintelligence

- Technological singularity
- Thinkism
- Transhumanism
- Robopocalypse